Dissonance suppression during harmonic mixing

MUS-17

Stefan Hamburger

July 5, 2017

Abstract

This work reviews the thesis *Techniques for automatic dissonance suppression in harmonic mixing* by Vittorio Maffei. Dance clubs require a continuous mix of music that does not interrupt the flow when switching between two songs. While traditionally done by a live DJ, new approaches can solve this problem fully or semi-automated. Maffei's thesis extends previous pitch-shifting approaches by modifying the spectral composition to remove dissonances.

Keywords: automatic mixing, Automated DJ, harmonic mixing, psychoacoustics, roughness.

Introduction

Ever since the rise of disco music during the 1970s and 80s, disk jockeys (DJs) are tasked with playing back music and keeping the crowd entertained. The end of a track is the most critical time because a bad follow-up track can clear the dancefloor when people take a break or leave the club. Therefore, good song selection and mixing is crucial for a DJ's career.

DJs quickly learned to adjust the tempo so that the beats from both tracks align (**beatmatching**), ensuring that people will not stop dancing during the transition. In addition, by only selecting follow-up tracks with a tempo and key similar to the current song, the mixing sounds more pleasant.

When computers revolutionized DJing around 2000, software could perform the beatmatching automatically. The now digital record collection can be sorted by both tempo and key, making it easy to find fitting tracks. New algorithms can change tempo and key independently, something that is impossible with physical turntables, which allows DJs to create more sophisticated mixes. While commercial DJ software can shift the key up or down, the state of the art in research, as described in Maffei's thesis [1], is not yet implemented in DJ software.

Automatic mixing is a controversial topic. On the one hand, veteran DJs consider it cheating because they think DJs should do the beatmatching themselves. On the other hand, the audience expectations have risen amd it becomes difficult to impossible for a DJ to perform all aspects of mixing by hand. For example, beatmatching is best done by a computer while song selection and interaction with the crowd are better done by a human. In the end, automatic mixing is just a tool and DJs can decide to what degree they want to use it.

Outside of clubs, automatic mixing can benefit all industries that require a constant stream of music, be they radio stations, music streaming services, video games, or retail environments.

Background

In music, two sounds are considered **harmonic** (also known as **consonant**) if they sound pleasant to a human listener. Harmony consists of two parts, chord progressions and the interactions between simultaneous tones. For harmonic mixing, we are only interested in the latter.

Harmony is universal to humans in all cultures, but it also depends on the personal listening experience. [3] With enough familiarity, what once

sounded dissonant can become harmonic, making it difficult to describe harmony generically.

However, certain intervals are considered to be more harmonic than others, this includes the unison, the octave, the perfect fifth and the major third. In the Circle of Fifths, all 12 semitones from Western music are ordered so that harmonic tones are close to each other. By looking at the distance of two tones on the Circle of Fifths, you can find out if they are harmonious.

Roughness

A more scientific explanation for harmony can be found in psychoacoustics. While it is still an open question how our brain works, there are some theories based on how audio waves are perceived by the ears and how their signals are transmitted to the brain.

One important concept is **roughness**, the psychoacoustic term for dissonance. Hearing two frequencies that are close together will make the listener uncomfortable because there is a beating and it becomes difficult to differentiate between both frequencies. In psychoacoustic terms, these frequencies lie in the critical bandwidth (CBW) of each other.

In 1965, R. Plomp and W. J. M. Levelt measured the roughness perception of various frequencies, and found that all study participants considered the roughness to be highest at 0.25 of the CBW. Using that data, we can calculate the degree of dissonance $R(f_1, f_2)$ given two frequencies:

$$R(f_1, f_2) = max((e^1 \cdot \tfrac{y}{0.25} \cdot e^{-\frac{y}{0.25}})^2, 0) \in [0, 1],$$

$$\text{with } y = \frac{|f_2 - f_1|}{CBW(\frac{f_1+f_2}{2})} \text{ and}$$

$$CBW(f) = 25 + 75 \cdot (1 + 1.4 \cdot (\tfrac{f}{1000})^2)^{0.69}$$

To calculate the roughness of complex tones T_1 and T_2, we sum the amplitude-weighted roughness of all pairs of partials:

$$R(T_1, T_2) = \frac{\sum_{i \in T_1} \sum_{j \in T_2} a_i \cdot a_j \cdot R(f_i, f_j)}{\sum_{i \in T_1} \sum_{j \in T_2} a_i \cdot a_j} \in [0, 1],$$

where a_i is a partial's amplitude and f_i is a partial's frequency.

Harmonic series

While roughness explains why a semitone step is more dissonant than a whole tone step, it does not explain bigger dissonant intervals like the tritone and major seventh. Instead, those intervals can be explained by the harmonic series.

All natural sounds, be they a vibrating string or an air column, will produce sound waves both at the fundamental frequency f and its multiples ($2f$, $3f$, $4f$, ...), these are called the **harmonics** or **partials**. The intensity of each harmonic varies, creating the tone color or timbre of an instrument. For example, in open-ended wind instruments only the odd harmonics are audible.

A human listener will perceive $f \triangleq 2f \triangleq 4f$ etc. to be at the same tonal level, this is called octave equivalence. In most tunings, $3f \triangleq 1.5f$ is the perfect fifth and $5f \triangleq 1.25f$ is the major third. Those intervals sound harmonic because most of their harmonics fuse together, e.g. if we play f_1 and $f_2 = 1.5f_1$, then the third harmonic from f_1 meshes with the second harmonics from f_2.

The tritone is one semitone below a fifth. When playing a tritone (e.g. $f_1 = C_4$ and $f_2 = F\sharp_4$), there will be a clash between $3f_1$ and $2f_2$ because those frequencies lie at 0.25 of the CBW. By taking all harmonics into account, we can explain all dissonances with roughness, aside from the minor variances based on personal preference.

Previous approaches

The first approaches to harmonic mixing used **key estimation**. There are numerous approaches to this, like looking at the spectrum generated by a Fourier transform and guessing the key based on the strongest partials. Using the Circle of Fifths, you can find a key combination that sounds harmonically good, and pitch-shift the tracks to the correct keys.

This approach is included in current DJ software like Traktor Pro 2. To make the Circle of Fifth more approachable to DJs, all keys are assigned a number from 1-12, and if two tracks have the

same number (or are off by one), the DJ knows he can mix the tracks without causing dissonances (cf. Figure 1).

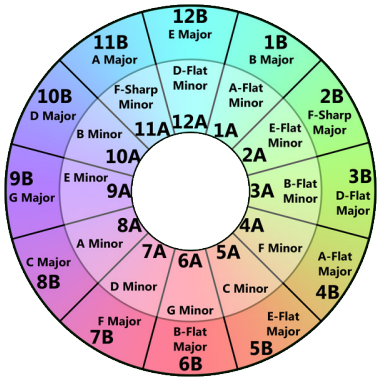

Figure 1: The EasyMix wheel, based on the Circle of Fifths. © Mark Davis, Camelot Sound.

The problem with key estimation is that even if the root key is detected correctly, we don't look at the melody or chord progression, which can cause dissonances between two tracks even if their key is identical. It will fail with atonal or chromatic tracks that are not composed using the major-minor tonality. Also, this approach is a dead end; the Circle of Fifths has existed for centuries and it is unlikely to advance from here.

Following that, **chroma-based mixing** was presented in 2014, based on research from 1999. In a chromagram, the frequencies from the spectrum are projected onto the 12 semitones (C, $C\sharp$, D, $D\sharp$ etc). This gives us more information to do the mixing; knowing the chord progression and melody, we can better synchronize the tracks. The main problem with this approach is the restriction to 12 semitones. Different tunings map a tone to different frequencies. Should there be an error and a tone is detected incorrectly, the mix will be off by a diminished second, the most dissonant interval.

Consonance-based mixing, presented in [2] in 2015, is the current state of the art. Unlike the previous two approaches, it is not based on music theory but on psychoacoustics. This makes sense because in the end, our mix should sound pleasant to a human listener, regardless of whether or not it follows music theory. Our ear can listen to a continuous frequency scale, not just the chromatic scale, so the mixing should be based on frequencies instead of semitones. Using the roughness measure, we can figure out which pitch-shift is the most consonant and shift by arbitrary amounts, not just by full semitones.

The new approach

The author of the thesis I am presenting finds that the consonance-based mixing is faulty in that it uses a very powerful measure to gather data (roughness), but it then ignores most of the data and does a simple pitch shift. Thanks to the advances in audio processing, it is now possible to do modify the spectral composition itself, not just by shifting it up or down but also by silencing certain frequencies. Therefore, the author builds on the system from [2] and adds dissonance suppression as a final step.

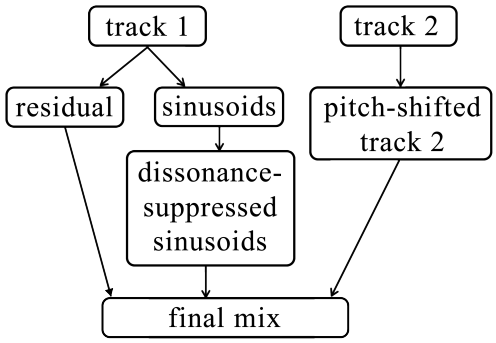

Figure 2: System overview

The system assumes that we already have two tracks (either from a Playlist Recommendation system or manual input) at a 44,100 Hz sampling rate. In a preprocessing step, both tracks are manually tempo adjusted to 120 bpm and shortened to 8 second fragments. That way, the system only needs to handle the harmonic mixing and can ignore beatmatching.

Next, the system performs a Short-time Fourier Transform (STFT) on both tracks, mapping from the time-domain waveform to a frequency domain spectrum. The author chose the Blackman window function, with a window size of 4096

samples and a hop size of 256 samples. This results in 1363 windows for each track. Each spectrum consists of 4096 bins, with a maximum frequency of 5000 Hz since any higher frequencies are not needed for harmonic mixing.

Once it has the spectral data, the system picks the 20 most prominent partials from track 1 and subtracts their waveform signal from the original waveform, resulting in a sinusoids and a residual part. The idea is that only the sinusoid part produces the dissonances and needs to be modified, while the residual part is left untouched to not introduce additional noise.

For the following steps, the $2 \cdot 1363$ windows are too much data for real-time processing, therefore the spectra are averaged onto 16^{th} notes, reducing the window data to only $2 \cdot 64$ time frames. The author found 16^{th} notes to be the best compromise between reducing data without adding too much noise.

For each of the 64 time frames, the system calculates the roughness between track 1 and the original track 2 plus 96 pitch-shifted versions of track 2 (from -6 to $+6$ semitones at $\frac{1}{8}^{th}$ semitone steps). By summing the roughness from all time frames, the system finds the pitch-shift that minimizes the roughness, and pitch-shifts track 2 accordingly. The pitch shift of a frequency f by the amount s can be calculated as follows:

$$f' = 2^{log_2(f) + \frac{s-48}{96}}, \text{ with } s \in [0, 96] \cap \mathbb{Z}$$

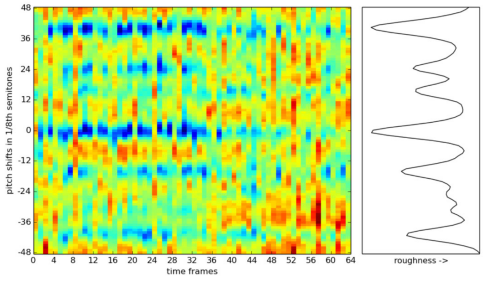

Figure 3: Pitch-shift selection, from [2].

Lastly, the system does a dissonance suppression, this is the new research that has not been described in any previous paper. The idea is to suppress certain parts of track 1 that produce the most dissonances. First, it uses the roughness values calculated above to select the time frames with the most roughness. Since the roughness numbers will vary between the tracks, it proved impossible to select an absolute cut-off value. Instead, the author decided on a percentile approach where the system chooses the time frames whose roughness would e.g. be in the 90^{th} percentile.

Once the roughest time frames are selected, the system would silence the sinusoid track during those time frames. This however caused a noticeable drop in volume and the resulting mix was not very pleasant. Therefore, the author went further and instead of silencing the whole sinusoid track, he looked at which partials contribute most to the roughness; here he again used a percentile approach. In addition to the partials of frequency f, he added the partials up and down one octave $(0.5f,\ 2f)$. Because of their lower amplitude, they would usually not end up being automatically selected but they cause the same amount of dissonance. Once he has a selection of rough partials, he silences those by -30dB.

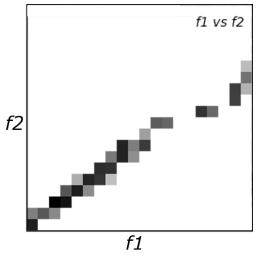

Figure 4: Roughness matrix of the partials from track 1 plotted against the partials from track 2. The darker a shade, the higher the roughness between those two partials. Taken from [1].

In the final step, all three audio files are merged to produce the final synthesized mix: the dissonance-suppressed sinusoid part from track 1, the residual part from track 1, and the pitch-shifted version of track 2.

Results

The author asked 13 musically-trained listeners to rate the consonance and pleasantness of mixes

produced by the system described in [2] versus the new dissonance suppression. It turns out that on mixes with a high roughness, the dissonance suppression resulted in better ratings, while on mixes that already had a high consonance, the dissonance suppression did not improve the ratings further.

Criticism

I do find many points of the paper problematic. First of all, the author has not provided any audio samples or code, which makes it hard to reproduce the results. In addition, his system is very similar to the system presented in [2] and most of his images were taken from that paper, making it difficult to rate his contribution to the research field.

Since the testing was only done on 8 second fragments, we cannot generalize the results; for longer samples the perceived consonance may be different. Also, I'd argue that if it takes musically-trained listeners in a controlled environment to rate the pleasantness of a mix, then people in a club, potentially drunk or drugged, will not notice much of a difference, rendering the usefulness of harmonic mixing moot.

Finally, there are many typos, which are at times confusing and always annoying (e.g. on page 47, "toneless" should read "tonalness"), but there are also technical errors. In section 4.5 (Optimal pitch-shift computation), the author writes "Once the phase analysis is completed," yet he never before mentioned a phase analysis, so it is unclear what he means by that. Formula 3.8 only works for the roughness of a single complex tone, but we actually need to look at the roughness between two complex tones. Also, the author recalculates the roughness during the partials suppression, even though it should be possible to reuse those values from the optimal pitch-shift computation.

Conclusion

The author admits that his chosen parameters may not be optimal, and that the results can be improved by picking different values. However, he does not hint at future research topics for harmonic mixing.

I find two topics of interest for the future. First of all, machine learning should be investigated. Given how machine learning has outperformed previous approaches in nearly all research fields, e.g. natural language processing, the same will likely be true for harmonic mixing.

Also, the current system makes no mention of volume adjustment. Orchestral tracks, e.g. from soundtracks, have huge variance in loudness. When listening to tracks from multiple soundtracks in shuffle play, the track transitions can cause huge disruptions, e.g. when a quiet track ends and an action track starts. Ideally, a mixing system should take this into account and quiet the action tracks while boosting calmer tracks. Though I admit that in dance music, this is not that big of a problem.

References

[1] V. Maffei. *Techniques for automatic dissonance suppression in harmonic mixing*. Master thesis at Politecnico di Milano, April 28th, 2016.

[2] R. Gebhardt et al. *Harmonic mixing based on roughness and pitch commonality*. Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15). Norway, 2015.

[3] Howard Goodall. *How Music Works*. 4-part TV series on Channel 4 (UK), 2006.