NEW SONORITIES FOR EARLY JAZZ RECORDINGS USING SOUND SOURCE SEPARATION AND AUTOMATIC MIXING TOOLS

Daniel Matz
University of Applied Sciences
Düsseldorf, Germany
daniel.matz@hotmail.de

Estefanía Cano
Fraunhofer IDMT
Ilmenau, Germany
cano@idmt.fhg.de

Jakob Abeßer
Fraunhofer IDMT
Ilmenau, Germany
abr@idmt.fhg.de

ABSTRACT

In this paper, a framework for automatic mixing of early jazz recordings is presented. In particular, we propose the use of sound source separation techniques as a preprocessing step of the mixing process. In addition to an initial solo and accompaniment separation step, the proposed mixing framework is composed of six processing blocks: harmonic-percussive separation (HPS), cross-adaptive multi-track scaling (CAMTS), cross-adaptive equalizer (CAEQ), cross-adaptive dynamic spectral panning (CADSP), automatic excitation (AE), and time-frequency selective panning (TFSP). The effects of the different processing steps in the final quality of the mix are evaluated through a listening test procedure. The results show that the desired quality improvements in terms of sound balance, transparency, stereo impression, timbre, and overall impression can be achieved with the proposed framework.

1. INTRODUCTION

When early jazz recordings are analyzed from a modern audio engineering perspective, clear stylistic differences can be identified with respect to modern recording techniques. These characteristics mainly evidence the technological and stylistic differences between the two eras. For example, solo instruments such as the saxophone or the trumpet often completely dominate the audio mix in early jazz recordings. At the same time, the rhythm section, i.e., double bass, piano, drums, and percussion, often falls in a secondary place, recorded or mixed with much lower intensity and often perceived as unclear and undifferentiated. Additionally, from today's perspective, early jazz recordings often present an unusual stereo image. Instrument groups are sometimes assigned to extreme stereo positions which can cause the solo instrument to be panned to the left and the accompaniment band panned to the right. As a consequence, the energy distribution over the stereo width is unbalanced and is often perceived today as irritating and disturbing, especially when listened through headphones.

Several initiatives have arisen that attempt to give such early recordings a more modern sonority. Remastering and Automatic Mixing (AM) techniques offer various methods for a sonic redesign of such recordings. However, given that the original individual stems of the instruments in the recordings are usually not available, these techniques can only achieve minor modifications to the sound characteristics of mono and stereo mixtures. In-depth remixing usually requires the original multi-track recordings to be available. For this purpose, sound source separation methods developed in the Music Information Retrieval (MIR) community can be useful tools to retrieve individual instruments from a given mix.

2. GOALS

The main goal of this study is to identify suitable signal processing methods to modify the above-mentioned characteristics in a selection of early jazz recordings. These methods are combined in a fully automatic mixing framework. In particular, we focus on modifying the audio mix in terms of transparency, stereo impression, frequency response, and acoustic balance in order to improve the overall perception of sound and the quality of the mix with respect to the original recording.

Our main approach for remixing is to modify the characteristic of the backing track to make it more present in the mix. We also aim at improving the acoustical and spatial balance of th audio mix. The solo signal is balanced with respect to its loudness and spectral energy to minimize spectral masking as well as to improve its position in the stereo image.

3. RELATED WORK

In the field of automatic mixing, several approaches have been presented in the literature. In [1], a method is proposed to automatically adjust gain and equalizer parameters for multi-track recordings using a least-squares optimization. In [12] the idea of modifying the magnitude spectrogram of a signal towards a target spectrogram called *target mixing*, is presented. Other approaches for automatic mixing of multi-track recordings have incorporated machine learning algorithms to perform the mixing process [16, 17].

In [14] and [19], several cross-adaptive signal processing methods for automatic mixing such as source enhancer, panner, fader, equalizer, and polarity and time offset correction are proposed. These modules can be combined into a full mixing application. In [4], the authors propose a knowledge-engineered autonomous mixing system and propose to include expert knowledge within an automatic mixing system. The included audio effects are automatically controlled based on extracted low-level and high-level features such as musical genre, instrumentation, and the type of sound sources. The authors evaluated the system using short four bar audio signals with vocals, bass, guitar, keyboard, and other instruments.

Harmonic-percussive source separation was used as pre-processing step for manual remixing in [6], in particular to adjust the sound source levels of the signals. To the authors' best knowledge, a framework for automatic remixing that suits the requirements discussed in section 2 has not been proposed so far.

4. PROPOSED METHOD

For our mixing framework, we propose the use of sound source separation techniques as a pre-processing step of the mixing process. For this purpose, we first isolate the solo instrument from the audio mix by applying an algorithm for pitch-informed solo and accompaniment separation [2]. The two separated signals, i.e., the *solo* and the *residual/backing* signal, are the starting point for the automatic remixing process. Additionally, based on the requirements discussed in section 2, our proposed framework comprises six subcomponents:

1. Harmonic-percussive separation (HPS)
2. Cross-adaptive multi-track scaling (CAMTS)
3. Cross-adaptive equalization (CAEQ)
4. Cross-adaptive dynamic spectral panning (CADSP)
5. Automatic excitation (AE)
6. Time-frequency selective panning (TFSP)

Figure 1 presents a block diagram of the proposed framework. There are three main signal pathways A, B, and C. If the CADSP is activated, pathway A is chosen. If CADSP is not activated, pathway B and C are chosen depending on whether the harmonic-percussive separation (HPS) is used. All signal paths output a stereo mix. In the following sections, the individual subcomponents are first described, followed by a description of the three proposed signal pathways.

4.1 Solo and Backing track Separation

The algorithm as proposed in [2] automatically extracts pitch sequences of the solo instrument and uses them as prior information in the separation scheme. In order to obtain more accurate spectral estimates of the solo instrument, the algorithm creates tone objects from the pitch sequences, and performs separation on a tone-by-tone basis. Tone segmentation allows more accurate modeling of the temporal evolution of the spectral parameters of the solo instrument. The algorithm performs an iterative search in the magnitude spectrogram in order to find the exact frequency locations of the different partials of the tone. A smoothness constraint is enforced on the temporal envelopes of each partial. In order to reduce interference from other sources caused by overlapping of spectral components in the time-frequency representation, a common amplitude modulation is required for the temporal envelopes of the partials. Additionally, a post-processing stage based on median filtering is used to reduce the interference from percussive instruments in the solo estimation.

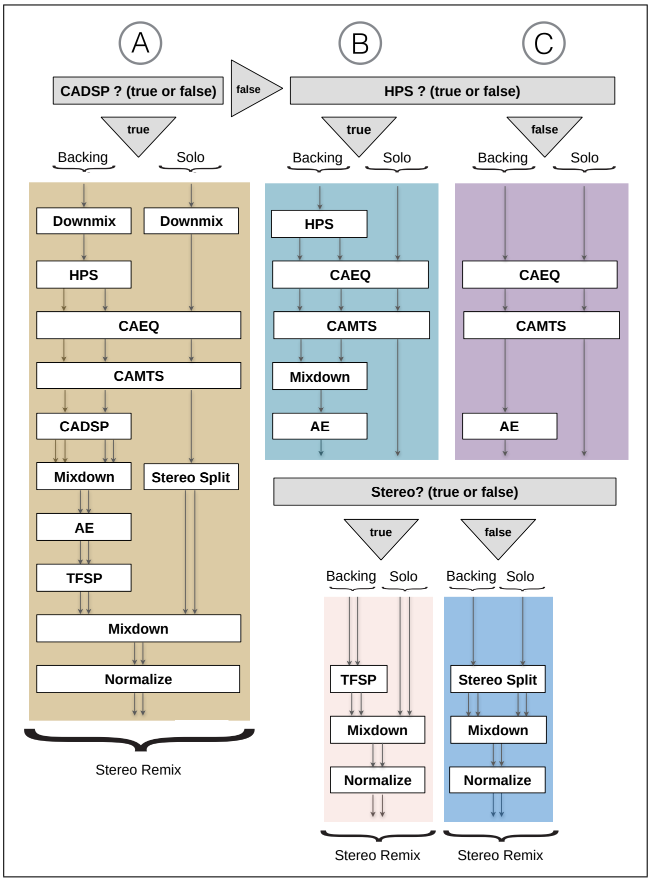

Figure 1: Signal flow-chart of the developed automatic remixing framework

4.2 Harmonic-percussive Separation (HPS)

We use the algorithm for harmonic-percussive separation proposed in [6]. The algorithm is based on median filtering of the magnitude spectrogram to split the original audio signal into its horizontal (harmonic sources) and vertical elements (percussive sources). In an automatic mixing context, these components can be understood as separate subgroups which can be processed individually and finally remixed.

4.3 Cross-adaptive Multi-track Scaling (CAMTS)

The method proposed in [19] which is commonly referred to as *automatic fader control*, is used for automatic scaling of the sound sources. The algorithm is used to automatically modify the amplification of separate sound sources. A psychoacoustic model based on the EBU R-128 standard [9] is used to compute the loudness of each track using a histogram-based approach. All tracks are individually amplified to be perceived as equally loud.

4.4 Cross-adaptive Equalizer (CAEQ)

We use the cross-adaptive equalizing algorithm proposed in [19] to obtain a spectrally balanced mixture. The main approach is to modify the spectral envelopes of the audio signals and to minimize the spectral masking between the solo signal and the backing track. The algorithm is a multi-band extensions of the CAMTS algorithm as discussed in section 4.3. The spectral characteristics of the separated signals are modified by enhancing or attenuating pre-defined frequency bands depending on the signal's perceived loudness with respect to the overall loudness. In contrast to the CAMTS algorithm, the loudness model proposed in [19] is used since it outperformed the loudness model based on EBU R-128 during informal testing. In particular, the mix results based on EBU-R 128 showed too strong of an emphasis on treble frequencies while lacking energy in the lower frequency range. We use a 10-band octave equalizer with second-order biquad IIR filters following [19] and frequency bands uniformly distributed over the audible frequency range. Standard frequency values based on [8] are used to adjust the center frequencies of the peak filter as well as the cutoff frequencies of the shelving filters.

4.5 Cross-adaptive Dynamic Spectral Panning (CADSP)

Dynamic spectral panning is a technique that allows the creation of a stereophonic impression in a given monophonic multi-track recording. We use the algorithm proposed in [15] to create a spatialization effect given multi-track signals. The method dynamically assigns time-frequency bins of the original tracks towards azimuth positions. The assignment reduces masking due to shared azimuth positions between multiple sound sources. This improves the overall transparency of an audio mix. In the cases where the original audio mix is a stereo track, it is first down-mixed to mono and then up-mixed to a new stereo image using the CADSP algorithm.

4.6 Automatic Exciter (AE)

The exciting algorithm improves the assertiveness of the backing track. The digital signal processing methods are implemented following the *APHEX Aural Exciter* described in [18]. The audibility of the mixed signal is enhanced by adding harmonic distortions in the upper frequency range. These distortions create additional harmonic signal components which improve the presence, clarity, and brightness of the audio signal.

The automation of the exciting step is implemented following a *target mixing* approach. Based on [5], the mixing parameters are iteratively adjusted to a *target energy ratio*. The target energy ratio is computed from the relationship between the energy of the high-pass filtered signal and the energy of the target signal. In the *side chain*, an asymmetric soft clipping characteristic, *harmonic generator block*, with adaptive threshold was used. This allows a level-independent distortion as well as the preservation of the signal dynamics [5].

4.7 Time-frequency selective Panning (TFSP)

Time-frequency selective panning improves the stereo image as well as the overall spatial impression of an audio mix. In our framework, the method for time-frequency selective panning presented in [3] was used. The azimuth positions of the sound sources are modified using a non-linear *warping function*. The stereo image is widened while the initial arrangement of the sound sources, as well as the sound quality of the original source is maintained. Within the proposed automatic remixing framework, the TFSP algorithm can be interpreted as an extension of the CADPS algorithm. The panning algorithm is only applied to the residual signal (see section 4.8.1). We set the aperture parameter ρ to a fixed value based on initial informal testing.

4.8 Processing Pathways

4.8.1 Signal path A (Main Path)

The main processing path includes all system components. Stereo files must be down-mixed to mono first due to constraints of the *cross-adaptive dynamic spectral panning* (CADSP) algorithm as detailed in section 4.5. All sound sources, which are initially distributed in the stereo panorama, are first centered to the mono channel and later redistributed over the stereo panorama again based on the harmonic-percussive sound separation [6]. This up-mixing step that can involve a modification of the stereo arrangement is only possible in this signal path.

The *cross-adaptive equalization* (CAEQ) and *multi-track scaling* (CAMTS) are the first processing steps in all three pathways. After applying the *dynamic spectral panning* (CADSP) to the percussive and harmonic signal components, all stereophonic signals are summed up to a backing track with a more homogeneous distribution of the sound sources. The backing track can now be processed with the *automatic excitation* (AE) and the *time-frequency selective panning* (TFSP) algorithms. The solo signal is split into stereo channels in the *Stereo Split* stage and scaled such that the overall gain remains constant. In the final *mix-down* step, the backing track is mixed with the solo track by adjusting the individual amplification factors as given by the CAMTS stage. If the cross-adaptive equalization (CAEQ) was performed, the spectral envelope of the backing track is perceivably modified due to the minimization of the spectral masking. The stereo sum signal is finally *normalized*.

4.8.2 Signal path B

Signal path B resembles signal path A, however, the equalization (CAEQ) and scaling (CAMTS) steps offer more ways to modify parameters due to the prior harmonic percussive separation stage.

4.8.3 Signal path C

In the signal path C, no harmonic-percussive separation is performed. The equalization (CAEQ) and scaling (CAMTS) are applied to both the backing and the solo track. However, the automatic excitation is only applied to the backing track since we particularly want to enhance the presence, clarity, and brightness of the backing track. As shown in figure 1, the time-frequency selective panning (TFSP) can only be applied to the backing track if it is a stereo signal. For monaural signals, the signal is split to the stereo channels (*Stereo Split*) and scaled such that the overall gain remains constant. Similar to signal path B, the signals are finally mixed down and normalized.

5. EVALUATION

5.1 Experimental Design

To evaluate the proposed framework, a listening test procedure was conducted following the guidelines of the *Multi Stimulus Test with Hidden Reference and Anchor* (MUSHRA) described in the ITU-R BS.1534-2 recommendation [11], and modifying them to fit the characteristics of this study. The main difference of our test with respect to the original MUSHRA is that a reference signal, which in our case would be an ideal mix of the original recording, is not available. Moreover, the notion of an ideal mix is ill-posed in the automatic remixing context.

The listening test was conducted in a quiet room and all signals were played using open headphones (AKG K 701). A total of 19 participants conducted the listening test. The participants included audio signal processing experts, professional audio engineers, music students (jazz, classical music), musicologists, as well as amateur musicians and regular music consumers. The average age of the participants was 30.7 years old. Further demographic information such as gender, hearing impairments, listening test experience, and educational background were also collected. A summary of the demographic information is presented in table 1.

The listening test was divided into five evaluation tasks, each focusing on a different subjective quality parameter. The following parameters were selected based on the ITU-R BS.1248-1 recommendation [10], and were adopted to our requirements: (QP1) Sound Balance, (QP2) Transparency, (QP3) Stereo/Spatial Impression, (QP4) Timbre, and (QP5) Overall Impression. In each evaluation task, a *training phase* was first conducted to allow the participants to familiarize themselves with the test material and to adjust playback levels to a comfortable one.

Gender	M	16
	F	3
Hearing impairment?	Yes	0
	No	19
Listening test experience?	Yes	9
	No	10
Expert in audio engineering?	Yes	11
	No	8
Educational background in music?	Yes	15
	No	4

Table 1: Demographics of the listening test participants

Following the training phase, an *evaluation phase* was conducted for each task. Five audio tracks as described in Table 2 with ten mixtures each were rated by the participants. The five tracks used in this study are part of the Jazzomat Database [1]. Among the presented mixtures, the original signal, eight mixes created with different configurations of the proposed framework, and an anchor signal (rhythm section reduced by 6 dB, the sum signal low-pass filtered at 3.5 kHz) were used. Table 3 gives an overview of all the remix configurations.

Title	Soloist (Instrument)	Style	Year
Body and Soul	Chu Berry (ts)	Swing	1938
Tenor Madness	Sonny Rollins (ts)	Hardbop	1956
Crazy Rhythm	J.J. Johnson (tb)	Bebop	1957
Bye Bye Blackbird	Ben Webster (ts)	Swing	1959
Adam's Apple	Wayne Shorter (ts)	Postbop	1966

Table 2: Dataset description

Mix	HPS	CAEQ	CAMTS	CADSP	AE	TFSP
1	off	on	off	off	on	off
2	off	off	on	off	on	off
3	off	on	on	off	on	off
4	on	on	off	off	on	off
5	on	off	on	off	on	off
6	on	off	off	on	on	on
7	on	on	on	on	on	on
8 (mono)	on	on	on	off	on	off

Table 3: Configurations of the eight remixes used in the listening test

The automatic exciting (AE) component is active in all the mixes. The panning (TFSP) algorithm is only activated in conjunction with the cross-adaptive dynamic spectral panning (CADSP). This way, a further stereo expansion of critical stereo recordings with an unbalanced stereo panorama is avoided. Mixture 8 was added to investigate the influence of the stereo effects (CADSP and TFSP) onto the input signals in the pre-processing step of pathway B that are mixed monophonic.

[1] A description of the Jazzomat Database is available at: `http://jazzomat.hfm-weimar.de/dbformat/dbcontent.html`

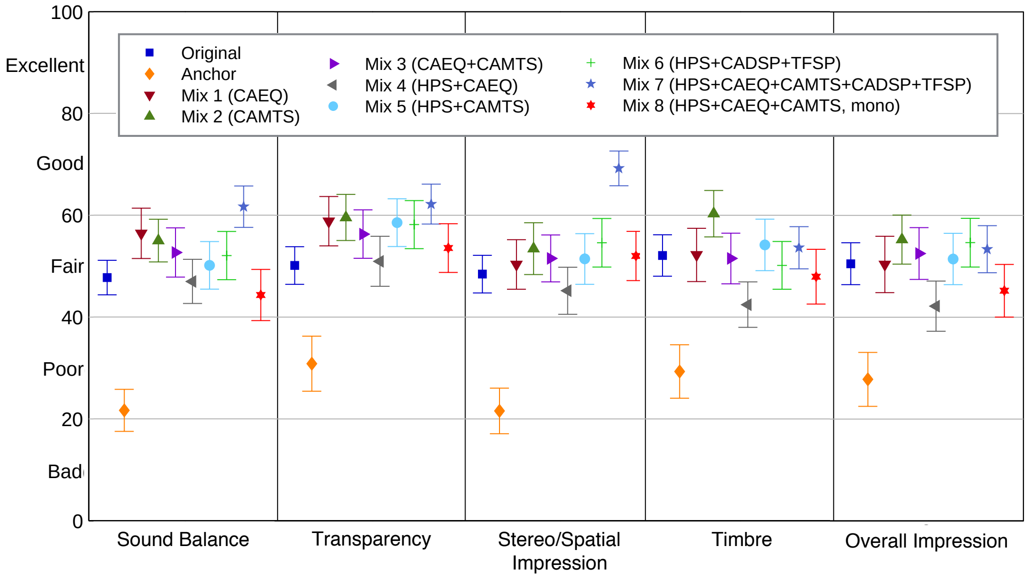

Figure 2: Listening test results for the five evaluated parameters.

5.2 Results

5.2.1 General

Figure 2 shows the results of the listening test for the five acoustic quality parameters. The figure legend summarizes all the system configurations that were evaluated. It is evident from the plot that the anchor stimulus was always correctly identified. Results also suggest that the use of harmonic-percussive separation does not bring perceptual quality gains for HPS+CAEQ (mix 4) when compared to the CAEQ (mix 3). Unexpectedly, results even got worse for the parameters timbre and overall impressions. Similarly, the combined settings in HPS+CAMTS (mix 5) do not show an improvement in the ratings when compared to CAMTS (mix 2).

To facilitate analysis of results, table 4 lists the percentage improvement obtained for each of the five quality parameters (QP), subject to the presence or absence of the individual framework components compared to the original signal. Mixes 4 and 5, which include the harmonic-percussive separation, were not listed due to the reasons previously described. The five mixtures listed in the table are further analyzed in the following sections.

	QP1	QP2	QP3	QP4	QP5
Mix 1 (CAEQ)	18 %	17 %	4 %	-	-
Mix 2 (CAMTS)	15 %	19 %	10 %	16 %	9 %
Mix 3 (CAEQ+CAMTS)	10 %	12 %	6 %	-	4 %
Mix 6 (HPS+CADSP+TFSP)	9 %	16 %	18 %	-	8 %
Mix 7 (All components)	29 %	24 %	43 %	3 %	6 %

Table 4: Percentage improvement of the remixed signal compared to the original audio recording subject to the presence (or absence) of the individual framework components shown for each of the five perceptual quality parameters.

5.2.2 Mix 1 (CAEQ)

Mix 1 does not include a prior separation of the residual component and outperforms the original mix for most of the quality parameters. The highest improvements were 18% for sound balance and 17% for transparency. However, for timbre and overall impressions, no improvement was observed.

5.2.3 Mix 2 (CAMTS)

Despite the absence of the harmonic percussive separation step, mix 2 showed improvements for transparency (19 %), sound balance (15%), and overall impression (9 %). The reason for the improvement in timbre by 16% is not entirely clear in this case; however, a possible explanation is that the increased loudness of the rhythm section led to more balanced dynamic levels and a clearer perception of the instrument and overall timbres.

5.2.4 Mix 3 (CAEQ+CAMTS)

The combination of the CAEQ and CAMTS components showed inferior results compared to the exclusive application of both components. However, the ratings are still slightly higher than the ratings of the original audio file.

5.2.5 Mix 6 and Mix 7

Both mixtures 6 and 7 outperformed the original audio file. The highest ratings were achieved with mixture 7 which was extracted with the full processing chain. In particular, the improvements compared to the original audio file were 29 % for sound balance, 24 % for transparency, as well as 43 % for stereo and spatial impression. The small improvements with respect to the overall impression are likely due to the individual aesthetic preferences of the listening test participants.

Additionally, to analyze the influence of the stereo effects to the input signals of pathway B (which are initially downmixed to mono), Table 5 presents the percentage improvement obtained with mix 7 (all components active) in comparison to mix 8 (mono).

QP 1	QP 2	QP 3	QP 4	QP 5
39 %	16 %	33 %	12 %	18 %

Table 5: Mean ratings of the five quality parameters for the additional usage of the stereo effects (CADSP+TFSP) in mix 7 compared with the non-processed monophonic input signal in the same framework setting of mix 8 (HPS, CAEQ, CAMTS, AE).

As can be observed in the table, the use of the CADSP and TFSP modules improved the ratings for all five quality parameters. The improvement was statistically significant for sound balance (39 %) and stereo/spatial impression (33 %).

6. CONCLUSIONS

In this paper, we proposed a prototype implementation of an *automatic remixing framework* for tonal optimization of early jazz recordings. The main focus was on improving the balance between the solo instrument and the rhythm section. The framework consists of six components which include different processing steps to modify the loudness, frequency response, timbre, and stereophonic perception of the separated sound sources. We compared different configurations of the framework and evaluated the improvement of the transparency of the backing track as well as the acoustic balance, stereophonic homogeneity, and overall quality perception. The evaluation was performed with a MUSHRA-like listening test based on the ratings given by 19 participants.

The usage of automatic equalization (CAEQ) and multitrack scaling (CAMTS) showed clear improvement in the quality parameter ratings, whereas the combination of both led to a smaller improvements than the independent application of each approach. The improvement based on harmonic-percussive separation (HPS) within the automatic mixing framework is not easy to assess. The usage of HPS in conjunction with CAEQ and CAMTS did not improve the ratings. On the other hand, HPS is a basic requirement for the application of CADSP on the backing track of mix 7, and therefore contributes to its consistent high ratings. HPS is irrelevant for the automatic excitation (AE) step, since it is applied to the full residual track.

Particularly with mix 7 (all components), the initially targeted improvements in sound balance, stereo and spatial impression, and transparency with respect to the original audio recording were achieved.

In future work, the most relevant processing modules must be further investigated and improved with respect to the aforementioned quality parameters. Modules that showed none or only minor improvements must be replaced and alternative algorithms must be evaluated for the given tasks. Promising algorithms seem to be a mastering equalizer [7] or dynamic range compression [13]. The additional use of semantic information of music genre and instrumentation seems to be another fruitful approach as discussed in section 3.

Finally, the integration of audio restoration methods such as denoising will likely help to remove unwanted background noise and spurious signals from the main signal to be processed.

7. REFERENCES

[1] Daniele Barchiesi and Joshua D. Reiss. Automatic target mixing using least-square optimization of gains and equalization settings. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009.

[2] Estefanía Cano, Gerald Schuller, and Christian Dittmar. Pitch-informed solo and accompaniment separation: towards its use in music education applications. *EURASIP Journal on Advances in Signal Processing*, 23:1–19, 2014.

[3] Maximo Cobos and Jose J. Lopez. Interactive enhancement of stereo recordings using time-frequency selective panning. In *Proceedings of the 40th AES International Conference on Spatial Audio*, Tokyo, Japan, 2010.

[4] Brecht De Man and Joshua D. Reiss. A semantic approach to autonomous mixing. *Journal of the Art of Record Production*, 8, 2013.

[5] Brecht De Man and Joshua D. Reiss. Adaptive control of amplitude distortion effects. In *Proceedings of the 53rd AES International Conference on Semantic Audio*, London, UK, 2014.

[6] Derry FitzGerald. Harmonic/percusssive separation using median filtering. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010.

[7] Marcel Hilsamer and Stefan Herzog. A statistical approach to automated offline dynamic processing in the audio mastering process. In *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, 2014.

[8] ISO International Organization for Standardization. Acoustics - preferred frequencies, August 1997.

[9] ITU Radiocommunication Bureau. Algorithms to measure audio programme loudness and true-peak audio level (rec. itu-r bs.1770-3), August 2012.

[10] ITU Radiocommunication Bureau. General methods for the subjective assessment of sound quality (rec. itu-r bs.1248-1), December 2003.

[11] ITU Radiocommunication Bureau. Method for the subjective assessment of intermediate quality level of audio systems (rec. itu-r bs.1534-2), June 2014.

[12] Zheng Ma, Joshua D. Reiss, and Black, Dawn A. A. Implementation of an intelligent equalization tool using yule-walker for music mixing and mastering. In *Proceedings of the 134th AES Convention*, Rome and Italy, 2013. AES.

[13] Stylianos-Ioannis Mimilakis, Konstantinos Drossos, Andreas Floros, and Dionysios Katerelos. Automated tonal balance enhancement for audio mastering applications. In *Proceedings of the 134th AES Convention*, Rome, Italy, 2013. AES.

[14] Enrique Perez Gonzales. *Advanced Automatic Mixing Tools for Music*. PhD thesis, Queen Mary University of London, London. UK, 30.09.2010.

[15] Pedro D. Pestana and Joshua D. Reiss. A cross-adaptive dynamic spectral panning technique. In *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, 2014.

[16] Jeffrey Scott and Youngmoo E. Kim. Analysis of accoustic features for automated multi-track mixing. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, USA, 2011.

[17] Jeffrey Scott, Matthew Prockup, Erik M. Schmidt, and Youngmoo E. Kim. Automatic multi-track mixing using linear dynamical systems. In *Proceedings of the 8th Sound and Music Computing Conference (SMC)*, Padova, Italy, 2011.

[18] Priyanka Shekar and Smith, III, Julius O. Modeling the harmonic exciter. In *Proceedings of the 135th AES Convention*, New York, USA, 2013.

[19] U. Zölzer. *DAFX: Digital Audio Effects*. John Wiley & Sons Ltd., second edition, 2011.