Audio Engineering Society

Convention Paper 8693

Presented at the 133rd Convention
2012 October 26–29 San Francisco, USA

Multi-track mixing using a model of loudness and partial loudness

Dominic Ward[1], Joshua D. Reiss[2], and Cham Athwal[1]

[1] *School of Digital Media Technology, Birmingham City University*

[2] *Centre for Digital Music, Queen Mary University of London*

Correspondence should be addressed to Dominic Ward (dominic.ward@bcu.ac.uk)

ABSTRACT

A method for generating a mix of multi-track recordings using an auditory model has been developed. The proposed method is based on the concept that a balanced mix is one in which the loudness of all instruments are equal. A sophisticated psychoacoustic loudness model is used to measure the loudness of each track both in quiet and when mixed with any combination of the remaining tracks. Such measures are used to control the track gains in a time-varying manner. Finally we demonstrate how model predictions of partial loudness can be used to counteract energetic masking for any track, allowing the user to achieve better channel intelligibility in complex music mixtures.

1. INTRODUCTION

A fundamental procedure in music mixing is the fader balance. The objective here is to scale the amplitude of each signal such that each sound source is audible, contributing to the overall loudness of the mixture. Automatic mixing techniques have been developed to achieve this goal with minimal or no human interaction, using amplitude averaging to derive the control signal for adaptive microphone gating [1] or the setting the channel gains directly [2].

More recent automix approaches can be split in two; those that employ machine learning techniques [3, 4, 5] and those founded on the cross-adaptive effect [6, 7]. In the latter case, where the focus of this paper lies, features are extracted from each of the input signals, relationships are analysed, and multiple signal processing decisions are formulated to enhance the mix (see [8] for an overview of multi-channel mixing systems).

Loudness features in particular are gaining interest

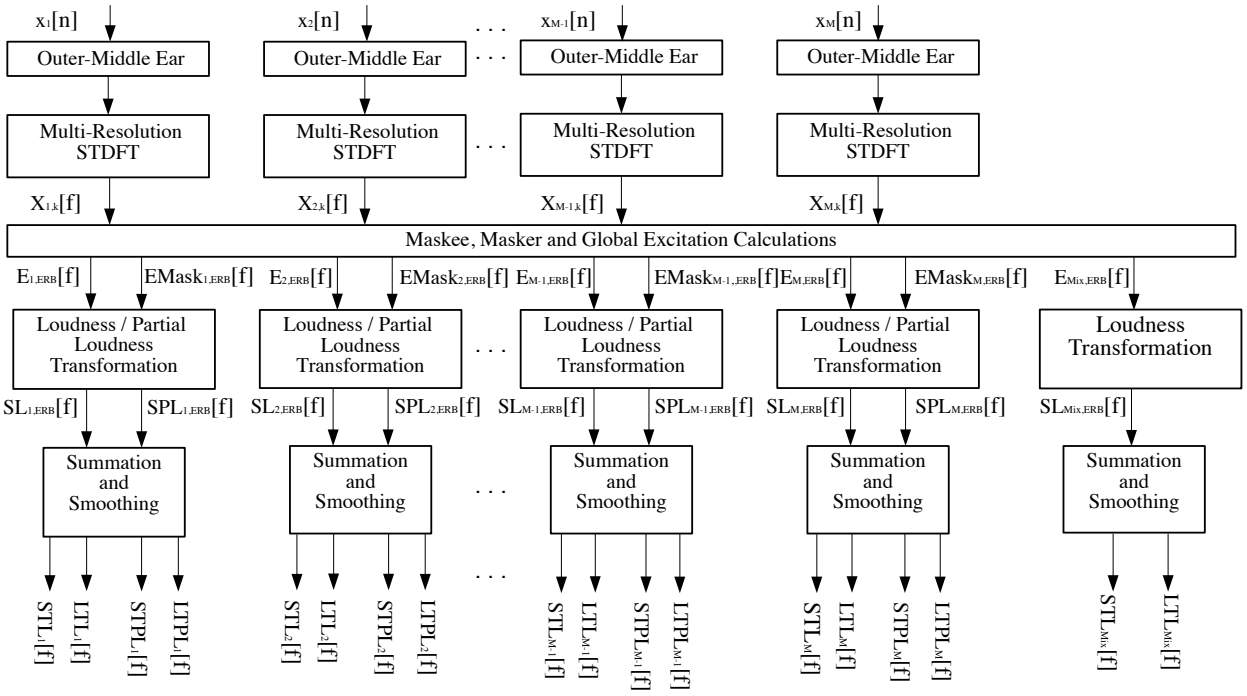

Figure 1: Block diagram of multi-channel loudness model for M input signals.

amongst the auto-mix community. For example, the system developed in [6] adapts the channel gains based on a time-varying average loudness. All input signals are filtered according to the shape of the ISO 226 loudness curves [9], optimising the channel gains such that a common loudness between the channels is maintained. Mansbridge et al. [7] developed an improved efficient implementation of the autonomous faders, using the EBU R-128 recommendation [10] as the basis of the loudness feature. The idea behind these systems is to mimic human operations by establishing control signals based on auditory perception. It is therefore imperative that the device employed to derive such controls accurately models the hearing system.

Although these previous methods have the advantage of being applicable to real-time performance, they are limited in that the loudness estimations are based on a simple frequency weighting. It is known that loudness depends on a multitude of variables, especially sound intensity, which can have a significant influence over the sonic impression of a mix when one considers playback level. The present work seeks to investigate the application of a sophisticated physiologically inspired loudness model to multi-track mixing. Briefly, the model used to perform the channel analysis in this paper is a multi-band model, whereas those used in previous works are single-band models (see [11] for a comparison). The model accounts for influences of intensity, frequency, bandwidth, duration and energetic masking on loudness perception. This has the advantage that both loudness in quiet and under conditions of masking can be quantified and used to inform the processing chain. The current system is designed for mixing a set of multi-track stems offline. A stem represents the combination of recordings corresponding to a single instrument e.g. an entire drum kit.

In this paper, we first detail the functionality of the selected loudness model. A procedure for achieving a time-varying common loudness between the inputs is then presented. Finally, a method for unmasking a prioritised track is detailed and initial results are discussed.

2. MODEL

Aichinger et al [12] combined the loudness models of Glasberg and Moore [13, 14] to perform a cross-analysis on musical instruments where each input may be masked by the combination of every other input. We have adapted the cross-analysis architecture presented in that paper to approach the task of automatic fader mixing. An overview of the multi-channel loudness model is depicted in Figure 1.

The loudness model accepts any number of channels and operates at typical sample rates. Calibration is essential and can be performed by measuring the sound pressure level of a 1kHz full scale tone at the listener's head, or eardrum if using headphones. All inputs to the model $x_m[n]$ are first passed through a 4097 coefficient FIR filter simulating the combined outer-middle ear magnitude response [13]. For testing purposes, the magnitude response of a pair of BeyerDynamic DT990 headphones placed on a GRAS artificial ear was measured. The FIR filter was redesigned to match the measured response combined with the middle ear transfer function as shown in Figure 2. A multi-resolution Short Time Discrete Fourier Transform (STDFT), comprising 6 parallel FFT's performs the spectral analysis. Each spectral frame $X_{m,k}[f]$ is filtered by a bank of level-dependent roex filters whose center frequencies range from 50Hz to 15kHz, the output of which yields the excitation pattern $E_{m,ERB}[f]$, where the frame number f is updated every millisecond. Such spectral filtering represents the displacement distribution and tuning characteristics across the human basilar membrane.

The excitation pattern is then transformed to a specific loudness pattern $SL_{m,ERB}[f]$ which represents the loudness at the output of each auditory filter. The summation of $SL_{m,ERB}[f]$ across the perceptual scale produces the total *unmasked* instantaneous loudness $IL_m[f]$.

To account for masking, each excitation pattern is recalculated as described in [15], along with an additional M excitation patterns required to formulate the background maskers for every channel, $EMask_{m,ERB}[f]$. The current implementation allows any combination of inputs to be used when generating the maskers. All excitation patterns are then transformed to a specific partial loudness pattern $SP_{m,ERB}[f]$ which describes loudness under inhibition [15]. This is summed to produce the total partial loudness $IPL_m[f]$.

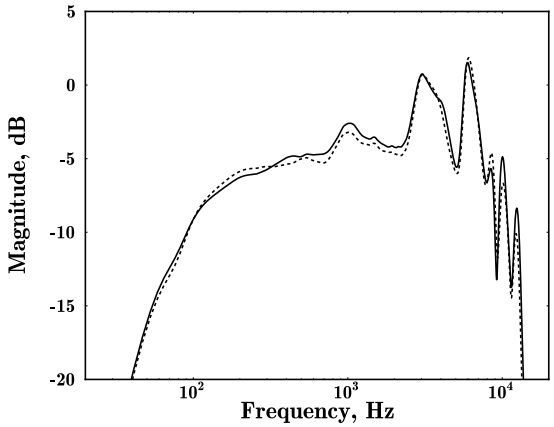

Figure 2: Combined headphone and middle ear response for left ear (—) and right ear (- - -).

Now that we have the unmasked and masked loudness values for every track, it would be useful to quantify the loudness of the entire mixture. Therefore, the final step is to calculate a single excitation pattern for the combination of all inputs, transformed and summed to yield the global loudness estimate $IL_{Mix}[f]$.

All of the above instantaneous loudness frames are smoothed by two separate temporal integration stages resulting in two perceptual measures; the short-term loudness $STL_m[f]$, describing the loudness perceived at any moment, and the long-term loudness $LTL_m[f]$, reflecting overall loudness judgements and memory effects [13]. Both the short-term partial loudness $STPL_m[f]$ and long-term partial loudness $LTPL_m[f]$ represent the same respective features, but under masked conditions. Thus, a total of four discrete time series are available at the output of the model per channel.

3. EQUAL LOUDNESS MIXING

In the present work, a mixture of M input tracks is described entirely by:

$$Mix[n] = \sum_{m=1}^{M} \alpha_m[n] x_m[n] \qquad (1)$$

Where α_m is the gain vector of the mth track. The aforementioned auto fader implementations focus on achieving a common average loudness per track by varying the track gains. The method described here also follows this objective. Establishing equal loudness amongst the channels ensures that no single channel heavily dominates the mixture, providing a well balanced sonic impression. In reality, multiple sources overlap in time and frequency, resulting in energetic masking. Consequently, as the number of tracks increase, instrument intelligibility decreases. This problem is dealt with in section 4 where an algorithm is presented to allow the user to selectively 'unmask' any given track.

3.1. Analysis Chain

In what follows, all inputs to the model are monophonic. During all analysis stages, the gains are applied to the power spectrum. Although gains adjustments could be applied to the waveforms, executing the multi-resolution STDFT only once reduces the computational cost when performing multiple loudness iterations. Furthermore, during the recursive procedures outlined below, spectral splatter is avoided which would have otherwise occurred when making modifications in the time-domain, heavily impacting the loudness measures. Once the gains have been solved for equal loudness, the input tracks are processed as described in section 3.3.2.

3.2. Loudness Normalization

In the same way that amplitude normalization achieves a common peak amplitude amongst the tracks, loudness normalization attempts to match the loudness of all tracks averaged over their entire duration. The goal of the proposed normalization procedure is to find the time-*invariant* gains of each track, such that the corresponding average loudness L_m contributes equally to a target average loudness $Lavg$. In this case $Lavg$ is given by:

$$Lavg = \frac{1}{M}\sum_{m=1}^{M} L_m \tag{2}$$

L_m may be formed by taking the mean $LTL_m[f]$ or $STL_m[f]$ over all frames where the channel loudness exceeds the value corresponding to absolute threshold, 0.003 sones. It is important to highlight that the original long-term loudness release time-constant given in [13] proves problematic when attenuating levels to a target loudness, leading to undershoot. Thus in this present work, we use the revised value of 200ms as suggested in [16]. The loudness values are then converted to loudness level in phons, and the track gain is given by:

$$\alpha_m[\cdot] = 10^{\frac{Lavg - L_m}{10}} \tag{3}$$

This procedure is repeated recursively until all channels converge to $Lavg \pm 0.5$ phons or after N iterations have occurred.

In order to maximise signal levels whilst allowing for further gain modifications, the above iterative method maintains a desired peak headroom HR within a given tolerance. Thus after all loudness values converge, the peak value of the mixture is measured and used to rescale the track gains. The loudness normalization routine is then reinitiated using a new $Lavg$ value measured after rescaling the levels. An overview of this procedure is given in Figure 3.

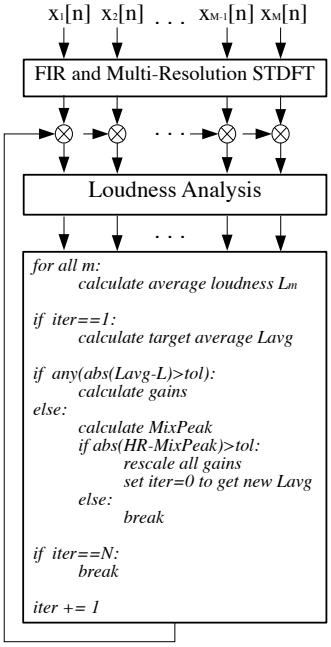

Figure 3: Loudness normalization. Pseudocode represents cross-adaptive feature processing.

3.3. Continuous Adjustments

Initial loudness normalization achieves a common average loudness between all channels over their entire audible duration. This is analogous to the way in which an audio engineer configures the faders when setting a rough mix. In the context of post production, volume automation is often applied to the multi-track recordings to attenuate intense instruments or amplify quieter instruments. This can be very time consuming, often requiring the user to manually program complex automation curves by hand. The proposed solution is to divide each track into short rectangular segments and perform the loudness normalization scheme per segment.

3.3.1. Segmentation

Segment detection is performed over each track to identify salient sustained periods of musical activity. A simple but fairly robust detection algorithm can be established by tuning the time-constants of the one-pole filter used to derive $LTL_m[f]$, and using this as a control signal, $Lc_m[f]$. If $Lc_m[f] > Mean(Lc_m[f]) \cdot k$, where k is a sensitivity parameter, channel m is said to be active. Averaging is performed over all frames where $Lc_m[f] > 0.003$ sones. Setting $k = 2/3$ provided a good estimate of loudness activity. Examples of segmentation performed on bass and vocal recordings are shown in Figure 4.

Performing logical operations on the boolean indices per track allows for the quantification of concurrent instrumentation, $present[f]$. This new vector stores the number of present channels per frame and is used to establish the frame locations where a gain change is permitted i.e. $present[f] > 1$.

After segmenting each track and locating overlapping audible portions, a target time-varying average loudness is formed as proposed in [7]. Accordingly, $Lavg[f]$ is equal to the sum of the channel loudnesses divided by $present[f]$. For each track, loudness normalization is performed on every segment over frames where $present[f] > 1$. If the channel count changes during this period, a new segment is formed and normalized. For each overlapping segment, $Lavg[f]$ is averaged and used to form the target loudness per track (in phons) for that segment.

For each track, all level adjustments (in decibels) and corresponding frame numbers are stored in $G_{f,m}$. An option has been included to decompose each of the longer segments into shorter windows, allowing finer gain adjustments to take place. This may be applicable when tracks are simultaneously active for long durations. Furthermore, a duration constraint of 2 seconds is placed on the normalization routine such that small overlapping segments are ignored.

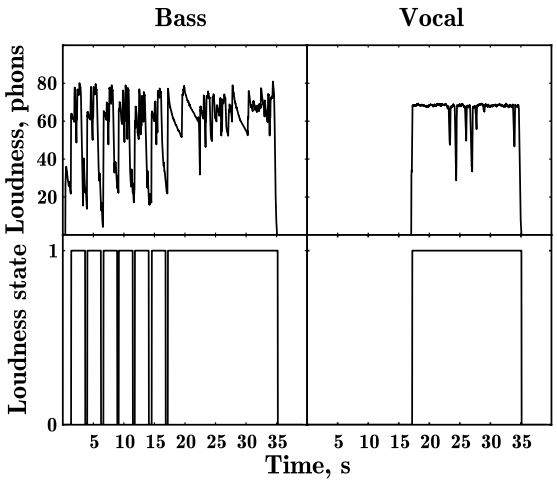

Figure 4: STL (top) and segmentation (bottom) of bass and male vocal recordings. Both recordings were loudness normalized.

3.3.2. Applying the gains

After normalizing the loudness of each segment, we are left with the gains $G_{f,m}$, where f and m denote the frame of each level adjustment and the track to be processed respectively. After converting each f to the corresponding sample index and performing linear interpolation on the sample level, the gains are smoothed according to:

$$\alpha_m[n] = (1 - \beta)G_m[n] + \beta\alpha_m[n - 1] \qquad (4)$$

where β is a smoothing coefficient. Finally, the gains are converted from logarithmic to linear values and applied to the original tracks in the time-domain. Figure 5 shows results of the global loudness and segmentation normalization procedures performed on 3 pure tones whose steady state portions have been staggered in time. Both the frequency and steady

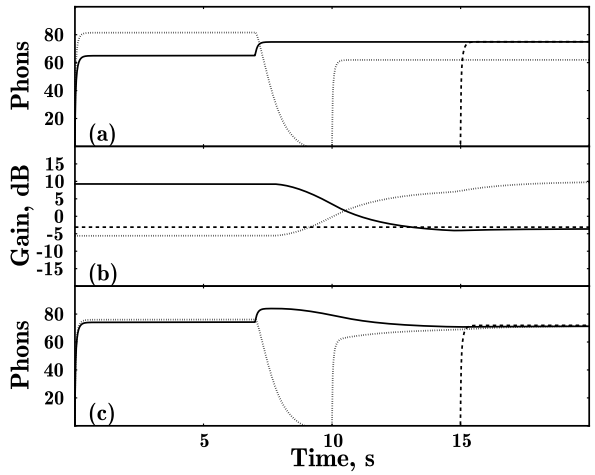

Figure 5: (a) LTL of 942Hz (—), 1035Hz (- - -) and 1325Hz (· · ·) sinusoids after initial loudness normalization. (b) Gains obtained from segmentation procedure. (c) LTL of same pure tones after loudness normalization and application of gains.

states levels were generated at random. The average long-term loudness was used as the comparative loudness measure. In subplot (c) it can be seen that the long-term loudness of all pure tones are approximately equal during periods of overlap. Note that the onset of the 1035Hz tone triggers a new gain estimation at time 15 seconds. By maintaining static gain values over the detected steady portions and smoothing between gain transitions, the original loudness trajectories are left intact.

4. LOUDNESS RESTORATION

Energetic masking occurs when the energy across the basilar membrane (the excitation pattern) of two or more sounds overlap causing a partial or complete reduction in loudness of each sound [17, 18, 19]. The partial masking of one sound by another is often explained in terms of the amount of effective overlap between the excitation patterns [20].

In musical performance, multiple instruments overlap in time and frequency, reducing the intelligibility of the individual instruments due to masking. It is imperative that the clarity of salient instrumentation is maintained in a mix. When conducting the fader balance, one option is to simply attenuate all instruments relative to the prioritised channel, reducing masking. Another is to increase the level of the masked channel, but this can lead to undesirable results such as masking of other instruments, clipping or an excessively loud signal. A more sensible approach is to attenuate only those channels that contribute to masking [21]. This section describes an approach to meet this criteria and is inspired by the perceptual speech reinforcement algorithm of Shin and Kim [22].

4.1. Partial Loudness

The excitation to specific partial loudness curves describe loudness reduction as a function of excitation level for a signal in noise [15]. As the excitation level exceeds masked threshold, the partial loudness rapidly increases until it approaches the loudness of the signal in quiet. In the current implementation the effect of each masker on a prioritised track is treated individually. Denote $PL_{p,m}$ as the average partial loudness of a selected prioritised track x_p when masked by a single source x_m where $m \neq p$. It is possible to identify any masking track and its influence on $PL_{p,m}$ by adjusting α_m and observing $\Delta PL_{p,m}$. When there is zero overlap between the two excitation patterns $E_{p,ERB}$ and $E_{m,ERB}$, the partial loudness is equal to its unmasked loudness value.

4.2. Procedure

After processing each track as described in section 3.3.2, the following method is carried out. First, for each track x_m where $m \neq p$, all segments overlapping with x_p are treated as potential maskers. Second, on a track by track basis, each segment is attenuated by 1dB. Both L_p and $PL_{p,m}$ are measured by averaging $STL_p[f]$ and $STPL_p[f]$ over a given segment, and converted to loudness level. Values below absolute threshold are ignored when averaging. Decrements of 1dB continue until $L_p - PL_{p,m} < 1$. This iterative process continues for each overlapping segment of every track.

In order to ensure all tracks are audible, a constraint is placed on the track gains. The current implementation forbids further level attenuation if the average partial loudness of any track within the *entire* mixture $PL_{m,mix}$ (in sones) falls below 0.003 or half its initial value. This latter constraint is used to avoid the tracks becoming too quiet relative to x_p. If x_p

becomes inactive, the track gains smoothly advance to 0dB, returning to the previously established equal loudness mix.

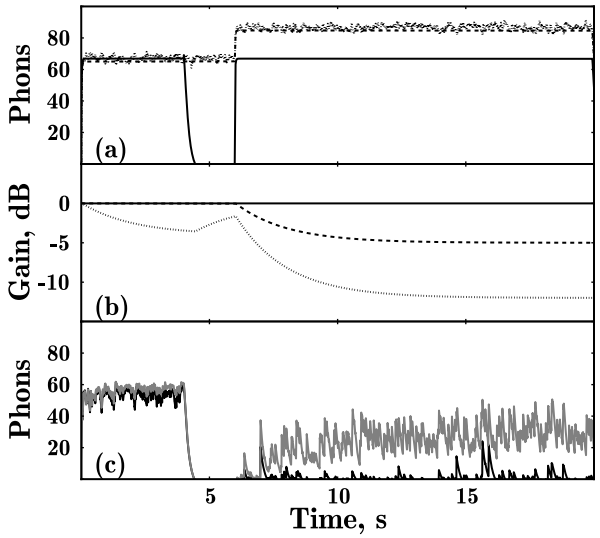

Figure 6: (a) STL of 1kHz sinusoid (—), 100Hz wide noise centered at 1kHz (· · ·) and 500Hz sinusoid (- - -). (b) Gains obtained after performing loudness restoration on the pure tone. (c) STPL of 1kHz tone before (black) and after loudness restoration (grey).

Figure 6 demonstrates the loudness restoration procedure applied to a 1kHz tone (maskee) at 70dB SPL presented in the combination of noise and a 500Hz tone (maskers). At time 6 seconds, both maskers were increased by 20dB. As shown by the gain trajectories in subplot (b), the amount of attenuation is dependent on both frequency content and level of the masker. Initially the 500Hz tone has little impact on the partial loudness of the 1kHz tone. As the 500Hz tone increases in level, an upward spread of masking towards higher frequencies occurs, and so the system begins to attenuate. Note that as the 1kHz tone becomes inactive, the noise begins to increase towards 0dB and hence return to its original loudness (time 5-6 seconds). Subplot (c) depicts the partial loudness of the 1kHz tone before and after loudness restoration. Complete loudness restoration is not achieved due to the gain constraint placed on both maskers in order to preserve their audibility.

5. PRELIMINARY EVALUATION

Two four-piece band recording sessions (one all electric, the other all acoustic) were used to test the system. The model was calibrated for headphone presentation, and an equal loudness mix per session was performed. After initial loudness normalization, the system applied few additional gain adjustments when running the segmentation procedure, and those that did occur were no more than ± 3dB. Admittedly, the dynamics of these recordings were all fairly stable over time. Allowing for finer gain adjustments seemed to maintain a constant loudness over time, especially for the vocal, suggesting that this option may indeed achieve a more successful loudness balance over the larger segment approach.

A comparison of relative loudness in quiet (no masking) revealed that for the electric rock band, the bass guitar was approximately twice as loud compared to any other instrument. This increase in loudness was less severe for the acoustic bass guitar but still noticeable. This indicates that the average long-term loudness may not be applicable to all instrumentation/playing styles, or perhaps the middle ear-transfer function over attenuates at low frequencies.

Loudness restoration was performed on the vocal recording within the electric session. The clarity of the vocalist was noticeably improved in the restored mix when compared to that of the straight loudness mix. As depicted in Figure 7, instrument attenuation is clearly dependent on the spectral proximity between each source and the prioritised track. Attenuation of the supplementary instruments is seen to be moderate due to the gain constraint used, but further unmasking could be achieved by parameter tuning. It is worth noting that severe masking can lead to audible 'ducking' as a result of all tracks periodically returning to 0dB during vocal inactivity. The gain contour of the guitar demonstrates this artefact. Longer time-constants when smoothing may aid to alleviate such impressions.

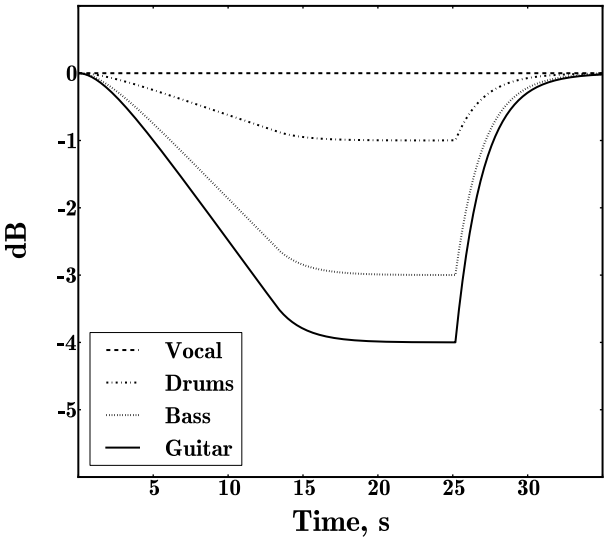

Figure 7: Obtained gain trajectories after performing loudness restoration on the lead vocalist, active during time 15-25 seconds.

6. CONCLUSIONS

A method for automatically mixing a set of multi-track recordings has been presented. The chosen method employs a sophisticated loudness model to perform the following operations:

- Decompose each track into sustained periods of musical activity
- Find the gains such that overlapping segments have a common overall loudness
- Restore the audibility of a selected track by reducing energetic masking

Initial subjective evaluations suggest that the loudness normalization routine performs well for most instruments, but tends to underestimate the loudness of low frequency instrumentation. The procedure advances previous automix approaches in that it considers the loudness of the instruments *within* the mixture. Thus, the audibility of a selected instrument can be restored, improving intelligibility, rather than relying solely on the method of equal loudness mixing.

A set of loudness matching experiments involving musical stimuli is planned to quantitatively assess the applicability of the model to a range of instrumentation. Further studies should also be conducted to assess how well model predictions of partial loudness generalises to musical signals, rather than 'laboratory stimuli' such as tones and noise. Aside from fundamental psychoacoustical testing, subjective quality tests should be carried out to evaluate the success of the system across genres. Finally, the current framework could be developed to allow the user to construct a hierarchical fader balance based on relative loudness values. This would allow for artistic demands such as 'instrument placement' within a mix that current automix implementations cannot accommodate.

7. REFERENCES

[1] D. Dugan. Automatic microphone mixing. *J. Audio Eng. Soc*, 23(6):442–449, July 1975.

[2] R. B. Dannenberg. An intelligent multi-track audio editor. volume 2, pages 89–94, San Francisco, August 2007. ICMC.

[3] B. Kolasinski. A framework for automatic mixing using timbral similarity measures and genetic optimization. In *Audio Engineering Society Convention 124*, May 2008.

[4] J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim. Automatic multi-track mixing using linear dynamical systems. In *SMC - 8th Sound and Music Computing Conference*, July 2011.

[5] J. Scott and Y. E. Kim. Analysis of acoustic features for automated multi-track mixing. In *International Society for Music Information Retrieval*, 2011.

[6] E. P. Gonzalez and J. D. Reiss. Automatic gain and fader control for live mixing. October, 2009. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE.

[7] S. Mansbridge, S. Finn, and J. D. Reiss. Implementation and evaluation of autonomous multi-track fader control. In *Audio Engineering Society Convention 132*. Audio Eng. Soc, April 2012.

[8] J. D. Reiss. Intelligent systems for mixing multichannel audio. In *17th International Conference on Digital Signal Processing, IEEE.*, July 2011.

[9] ISO. Normal equal-loudness-level contours. Technical Report 226, International Standard Organisation, 2003.

[10] EBU Recommendation R 128. Loudness normalisation and permitted maximum level of audio levels. Recommendation, EBU, August 2010.

[11] E. Skovenborg and S. H. Nielsen. Evaluation of different loudness models with music and speech material. In *Audio Engineering Society Convention 117*, August 2004.

[12] P. Aichinger., A. Sontacchi., and B. Schneider-Stickler. Describing the transparency of mixdowns: The masked-to-unmasked-ratio. In *Audio Engineering Society Convention 130*, May 2011.

[13] B. R. Glasberg and B. C. J. Moore. A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc*, 50(5), May 2002.

[14] B. R. Glasberg and B. C. J. Moore. Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds. *J. Audio Eng. Soc*, 53(10):906–918, October 2005.

[15] B. C. J. Moore, B. R. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc*, 45(4):224–240, April 1997.

[16] B. C. J. Moore, B. R. Glasberg, and M. A. Stone. Why are commercials so loud? - perception and modeling of the loudnes of amplitude-compressed speech. *J. Audio Eng. Soc*, 51(12):1123–1132, December 2003.

[17] B. Schauf. Partial masking. *Acustica*, 14:16–23, 1964.

[18] S. S. Stevens and M. Guirao. Loudness functions under inhibition. *Perception and Psychophysics*, 2(10):459–465, 1967.

[19] A. M. Richards. Monaural loudness functions under masking. *Journal of the Acoustical Society of America*, 44(2):599–605, March 1968.

[20] E. Zwicker and B. Scharf. A model of loudness summation. *Psychol. Rev.*, 72(1):3–26, 1965.

[21] E. P. Gonzalez and J. D. Reiss. Improved control for selective minimization of masking using interchannel dependancy effects. In *11th Int. Conferance on Digital Audio Effects*, September 2008.

[22] J. W. Shin and N. S. Kim. Perceptual reinforcement of speech signal based on partial specific loudness. *IEEE Signal Processing Letters*, 14:997–890, 2007.