AUTOMATIC TARGET MIXING USING LEAST-SQUARES OPTIMIZATION OF GAINS AND EQUALIZATION SETTINGS

Daniele Barchiesi

Centre for Digital Music,
Queen Mary, University of London
United Kingdom
ee08m168@elec.qmul.ac.uk

Josh Reiss

Centre for Digital Music,
Queen Mary, University of London
United Kingdom
josh.reiss@elec.qmul.ac.uk

ABSTRACT

The proposed automatic target mixing algorithm determines the gains and the equalization settings for the mixing of a multi-track recording using a least-squares optimization. These parameters are estimated using a single channel target mix, that is a signal which contains the same audio tracks as the multi-track recording, but that has been previously mixed using some unknown settings.
Several tests have been done in order to evaluate the performances of two different approaches to the optimization, namely the sub-band estimator and the FIR filters estimator. The results show that, using the latter technique, the proposed algorithm is able to retrieve the parameters originally applied to the target mix.
This achievement can be useful for remastering applications, where both the original recording sessions and the final mix are available, but there is the need to retrieve the mixing parameters originally applied to the various audio tracks.

1. INTRODUCTION

Automatic mixing is an emerging research field whose objective is to provide new tools for sound engineers in order to partially or completely automate the mixing process for both live and studio productions.
In particular, the goal of automatic target mixing is to derive the parameters in the mixing of a multi-track recording based on a target mix. The target can be, in theory, any audio signal which the user can choose as a reference because of some particular qualities like a certain equalization curve for a given instrument or a certain balance between the amplitude of the various tracks.
Although the algorithms shown in this paper can be applied to any kind of target, their evaluation in the context of a target mix and a multi-track recording which contain different source signals is not a trivial problem and is left for further research. In this work, we focus on the case where the target mix was obtained applying a set of unknown parameters to the multi-track recording. Thus, the value of the estimated parameters should depend only on the mixing process and not on the audio content of the tracks.

The main application of the proposed technique is remastering [1]. Figure 1 shows a common situation where an old analog multi-track session is remastered in a new digital format, applying modern techniques in order to improve the quality of the recording.
Firstly, the single tracks are converted to the digital domain through the ADC, then processing is applied on the signals in order to eliminate noise and other typical artifacts present in old media formats. At this point, the single tracks have to be mixed but, in most cases, the original mixing parameters are not available. Using our proposed automatic target mixing (ATM), it is possible to retrieve these parameters from the final mix of the analog multi-track recording. Optionally, the mixing process can include an up-mixing stage, which can be used to create multichannel versions of the recording.

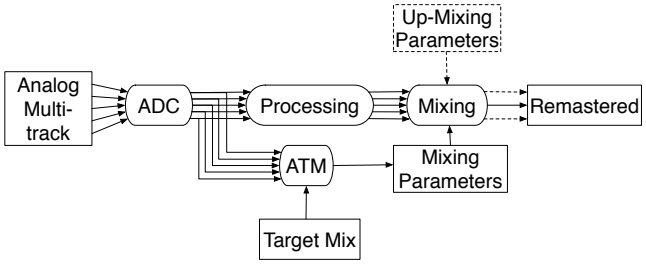

Figure 1: *Application of automatic target mixing: remastering of an analog recording.*

In addition to remastering, our algorithm can be useful since some artists have recently released multi-track recordings along with the final mix of some of their works [2], giving the listeners the opportunity to create their own mixes. From this prospective, the automatic target mixing can be a valid teaching instrument that shows how famous and skilled sound

engineers have mixed a certain song.

Some commercial products exist that suggest mixing parameters according to a target [3], [4]. They are based on a comparison between single tracks or whole mixes, and thus they do not define specific parameters for each one of the recordings in a multi-track session. Other previous work in automatic mixing based on a target has been done by Reed [5], who designed a system which suggests equalization settings depending on a users' goal (for instance make the sound "brighter", or make the frequency content more "smooth") and uses an inductive nearest neighbour machine learning algorithm.

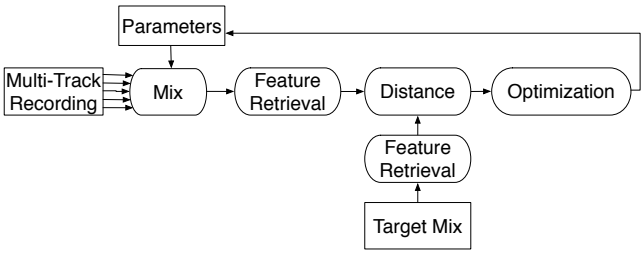

Figure 2: *Automatic Target Mixing Framework*

Kolasinski [6] proposed an automatic target mixing technique which has inspired our research and whose general framework is depicted in figure 2.
As can be seen, the multi-track recording is mixed using a certain set of parameters, then a feature or a set of features is extracted from the resulting mix and from the target mix. The distance between those features is computed and is input to an optimization algorithm. This estimates the set of parameters which minimize the distance from the target, so that this particular set can be used to build the estimated mix.

The algorithm designed by Kolasinski [6] aims to find the gains to apply to each track which minimize the euclidean distance between the spectrum histograms [7] of estimated and target mix. In his work, the optimization task is performed by a genetic algorithm [8]. This is a search technique that has been proved to be effective in locating the global minimum of large and uneven parameters' spaces. However, for the purpose of this application, the results are quite poor as the number of tracks increases and the algorithm is computationally expensive. Moreover, an iterative optimization like the genetic algorithm can only converge to an approximate solution of the problem.
In the next section a geometric approach is described in order to analytically solve the optimization problem.

2. SOLVING THE OPTIMIZATION PROBLEM. A GEOMETRIC APPROACH

Assume that the parameters of the mix that we want to estimate are the gains applied to each track x_i and that we are comparing a linear feature extracted from mix and target, that is a feature F such that

$$F(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 F(x_1) + \alpha_2 F(x_2) \quad \forall \alpha \in \mathbb{R} \quad (1)$$

Then the optimization task can be solved analytically using least squares.

Let $t = F(\text{target mix})$ be the feature extracted from the target mix and $v_i = F(x_i)$ the feature extracted from the i-th track of the multi-track recording.
t can be represented as a vector in an M-dimensional space (where M is the size of the feature), as is shown in figure 3. We can define λ as the sub-space generated by any linear combination of the vectors v_i with positive coefficients. The dimensionality of λ is equal to the number of tracks N and, in general, we have $N << M$. The mix which minimizes the distance from the target is the projection of t on the subspace λ.

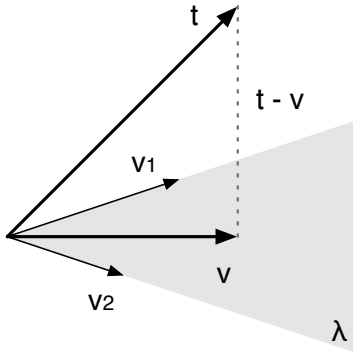

Figure 3: *Geometric Representation of Target Mix and Multi-track Recording*

We can write the projection vector v as a linear combination of the tracks vectors v_i with the gains $\hat{\alpha}$ that we want to retrieve.

$$v = \sum_{i=0}^{N-1} \hat{\alpha}_i v_i$$

A is defined as the matrix whose columns are the vectors v_i

$$A = \left(\begin{bmatrix} \\ v_1 \\ \\ \end{bmatrix} \quad \begin{bmatrix} \\ v_2 \\ \\ \end{bmatrix} \quad \dots \quad \begin{bmatrix} \\ v_N \\ \\ \end{bmatrix} \right)$$

and v is then written as

$$v = A\hat{\alpha}$$

The vector $t - v$ is orthogonal to each of the vectors v_i. Therefore the inner products between v_i and $t - v$ must be zero for each i. This leads to the following equation:

$$A^T(t - v) = 0$$

Substituting, we obtain

$$A^T t - A^T A\hat{\alpha} = 0$$

which implies

$$\hat{\alpha} = (A^T A)^{-1} A^T t \quad (2)$$

Equation 2 is known as the least squares method [9]. The only condition for the matrix $A^T A$ to be invertible is that the vectors v_i must be linearly independent, which is an assumption we can make in general.

As already mentioned, the least square optimization of a multi-track recording can be performed using any feature extracted from the audio signals as long as it respect equation 1. Unfortunately many features (for instance, those based, on the power spectrum or on psycho-acoustical models) are not linear and therefore cannot be used in this framework. However, for the purpose of our problem, we will only consider the signals in the time domain and their Fourier transform as the feature vectors.

In particular, choosing the latter, we can extend the least-squares approach to the equalization problem applying different gains to different frequency bands of the tracks. This results in increasing the dimension of the sub-space λ, choosing sets of orthogonal vectors $w_{i,j} \perp w_{i,k}$ such that

$$\sum_{j=1}^{P} w_{i,j} = v_i$$

where P is the number of frequency bands.
This algorithm is estimating the equalization curve applied to the various tracks in the target mix using a piecewise constant function for each track, and will be referred as the sub-band estimator. As P increases, the estimation is more accurate and, in theory, able to approximate any transfer function at the expense of a higher computational cost.

3. TARGET EQUALIZATION

As described in the previous section, the equalization parameters can be retrieved using the sub-band estimator. However, this method does not reflect the way equalization is usually performed, using combinations of FIR or IIR filters. There exist various techniques for matching a target transfer function to a filter's impulse responses [10],[11]. For instance, the method described by Lee [12] is based on a least squares optimization and is similar to the one proposed in this section. These techniques are often applied to loudspeaker and microphone equalization. Yet, to the authors' knowledge, filter optimization techniques have not been adapted to the target mixing problem, where a different transfer function is applied to each track.

The FIR filters estimation technique described in this section finds the coefficients of filters which, when applied to the tracks of the multi-track recording, minimize the Euclidean distance between the target mix and estimated mix.

3.1. Linear Predictive Model

Since the proposed technique is inspired by linear prediction, the basic theory of this model is presented, and then extended to the target equalization problem.

Let v be an audio signal, then its value $v(n)$ at time n can be estimated as a linear combination of its S previous samples:

$$\tilde{v}(n) = \sum_{j=0}^{S-1} \alpha_j v(n - j)$$

The goal of the linear predictive model is to find the coefficients $\hat{\alpha}_j$ that minimize the squared euclidean distance between signal and predicted signal. We can define this squared distance as a function of α

$$J(\alpha) = ||\tilde{v} - v||^2$$

and compute the coefficients $\hat{\alpha}$ setting the gradient $\nabla J(\alpha)$ to zero.
This results in solving the following system of linear equations:

$$R_l = \sum_{j=0}^{S-1} \hat{\alpha}_j R_{j-l} \quad (3)$$

where R_l is the autocorrelation of the signal v and R_{j-l} the j-shifted autocorrelation.
The coefficients $\hat{\alpha}$ define an FIR filter which is applied to the signal v, therefore we can use the same approach and define a new squared distance function in order to solve the target equalization problem.

3.2. FIR Coefficients Estimation

Let t be the target mix, M the length of the target, N the number of tracks in the multi-track recording and P the or-

der of the filter we want to estimate. Then we can define the following squared distance function:

$$J(\alpha) = ||t - \sum_{i=0}^{N-1} v_i * \alpha_i||^2 = \sum_{n=0}^{M+P-1} \Big[t(n) - \sum_{i=0}^{N-1}\sum_{j=0}^{P-1} \alpha_{ij} v_i(n-j)\Big]^2$$

The partial derivative is computed as:

$$\frac{\partial J(\alpha)}{\partial \alpha_{kl}} = 2 \sum_{n=0}^{M+P-1} \Big[t(n) - \sum_{i=0}^{N-1}\sum_{j=0}^{P-1} \alpha_{ij} v_i(n-j)\Big] v_k(n-l)$$

and then the gradient $\nabla J(\alpha)$ is set to zero. This corresponds to solving the following system of linear equations:

$$\sum_{n=0}^{M+P-1} t(n) v_k(n-l) = \sum_{i=0}^{N-1}\sum_{j=0}^{P-1} \hat{\alpha}_{ij} \sum_{n=0}^{M+P-1} v_i(n-j) v_k(n-l) \quad (4)$$

The first sum on the left side of equation 4 is the correlation $C_l(t, v_k)$ between the target t and the k-th track v_k, while the sum over n on the right side is the shifted correlation $C_{j-l}(v_i, v_k)$ between the i-th track and the k-th track. Equation 4 can be written as:

$$C_l(t, v_k) = \sum_{i=0}^{N-1}\sum_{j=0}^{P-1} \hat{\alpha}_{ij} C_{j-l}(v_i, v_k) \quad (5)$$

$$\forall\, k = 0, \ldots, N-1 \quad l = 0, \ldots, P-1$$

The correlation $C_l(v, w)$ between two vectors v and w is the inner product between v and the l-shifted version of the vector w. Thus, we can build a matrix A whose columns contain the tracks v along with the l-shifted tracks:

$$A = \left(\begin{bmatrix} v_1 \\ \\ \end{bmatrix} \quad \begin{bmatrix} v_1 \\ \\ \end{bmatrix} \quad \ldots \quad \begin{bmatrix} \\ \\ v_1 \end{bmatrix} \quad \ldots \quad \begin{bmatrix} v_N \\ \\ \end{bmatrix} \quad \ldots \quad \begin{bmatrix} \\ \\ v_N \end{bmatrix} \right)$$

and write equation 5 in matrix form:

$$(A^T A)\hat{\alpha} = A^T t$$

which can be solved again with the least square method:

$$\hat{\alpha} = (A^T A)^{-1} A^T t$$

It is interesting to consider a particular case of the equation 5. If we set the order of the filter P to one, then we are simply applying different gains to each track, and the matrix A is equal to the one defined in the section 2. Therefore, the FIR estimation can be viewed as a generalization of the geometric approach described to retrieve the gain settings. This least squares estimation minimizes the norm of the error function $J(\alpha)$ in the time domain. But, since the Fourier transform doesn't affect the norm of a function, up to a factor $1/2\pi$, then this means that the distance in the Fourier domain between target and estimated mix will also be minimized.

The only parameter we have to chose is P. As with the sub-band estimator, the number of parameters determines the accuracy of the algorithm, but also its computational cost. The next section describes some experiments that have been performed in order to compare the two methods.

4. ESTIMATORS EVALUATION

In general we know neither the number nor the type of filters applied to the multi-track recording and, therefore, the estimation of equalization parameters must be robust to different choices of filters. For this reason we have designed some experiments in order to compare the performances of the sub-band and FIR estimators.
The algorithms were tested on a 6-tracks recording that contains different instruments on each channel. The audio files are sampled at 16 bits/44100 Hz, and their duration is approximately 30 seconds.
As a preliminary result, we estimated parameters for the equalization of a single track then we applied our algorithm on the whole multi-track recording in order to retrieve different parameters for each channel.

Firstly we designed a 256-order FIR multiband filter, whose impulse response is depicted in figure 5 and whose frequency response is shown in figure 6. This filter can be viewed as a graphic equalizer with extreme settings where the frequency bands 0-0.2KHz, 1.5-3KHz and 15-22KHz have unitary gain and the other bands are fully attenuated.
The convolution of one instrument of our multi-track recording with this filter was the target mix. Then, we applied iteratively both the sub-band estimator and the FIR estimator, increasing progressively the number of estimation parameters. For each iteration, we computed the euclidean distance between target and estimated mix.

Figure 4 shows the distance as the number of parameters approaches the order of the filter. As can be seen, the sub-band estimation does not improve much as the number of parameters increases, while the distance function decreases to zero using the FIR estimator. The experimental results suggest that there must be a correspondence between the distance function and the energy present in the target's impulse re-

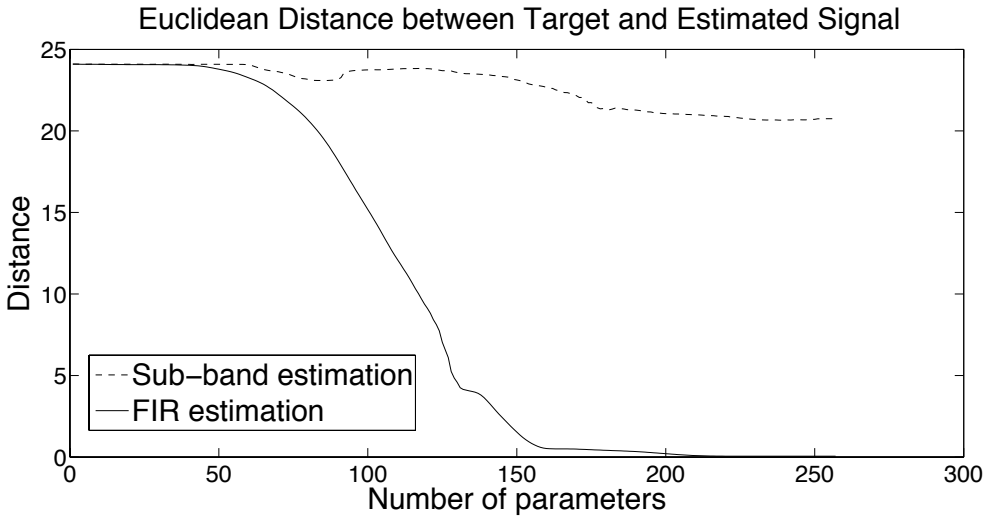

Figure 4: *Distance from target for a high order FIR multiband filter*

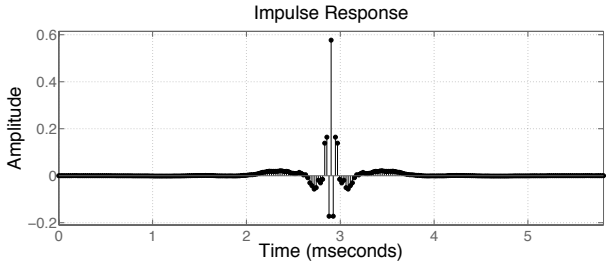

Figure 5: *Impulse response of a high order FIR multiband filter*

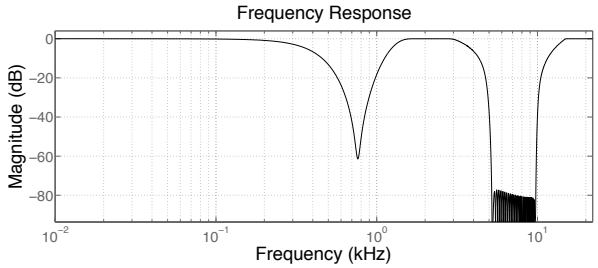

Figure 6: *Frequency response of a high order FIR multiband filter*

sponse. In fact, most of the energy of the impulse response depicted in figure 5 is present at less than 3 ms (corresponding to 130 samples) and we can see that, when the number of parameters of the FIR estimator reaches a value around 130, the distance drops to a very small value.
This behaviour has been confirmed using other target impulse responses. It implies that, as long as the target filter has an impulse response whose energy decays quickly enough, then the FIR estimator will produce a good approximation of it with a finite number of coefficients.

Figures 7 and 9 show the impulse and frequency response of an 8th order IIR low-pass filter used to test the estimators. The cutoff frequency is 1KHz and the design is based on a Butterworth analog prototype. The distance between the target and estimated signal as a function of the number of parameters is shown in figure 8.
Again we can see that the sub-band estimator doesn't produce good results, while using the FIR estimator the distance from target decreases towards zero as the number of parameters increases.
In this case, the FIR estimator can't reproduce the exact target impulse response because we would need an infinite number of parameters. However, since the energy of the target impulse response decays quickly to zero, we can achieve a good approximation even with a small number of coefficients.

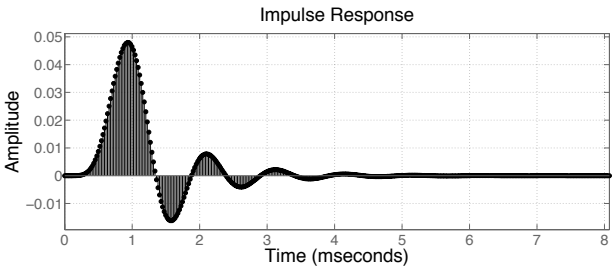

Figure 7: *Impulse response of an 8th order IIR lowpass filter*

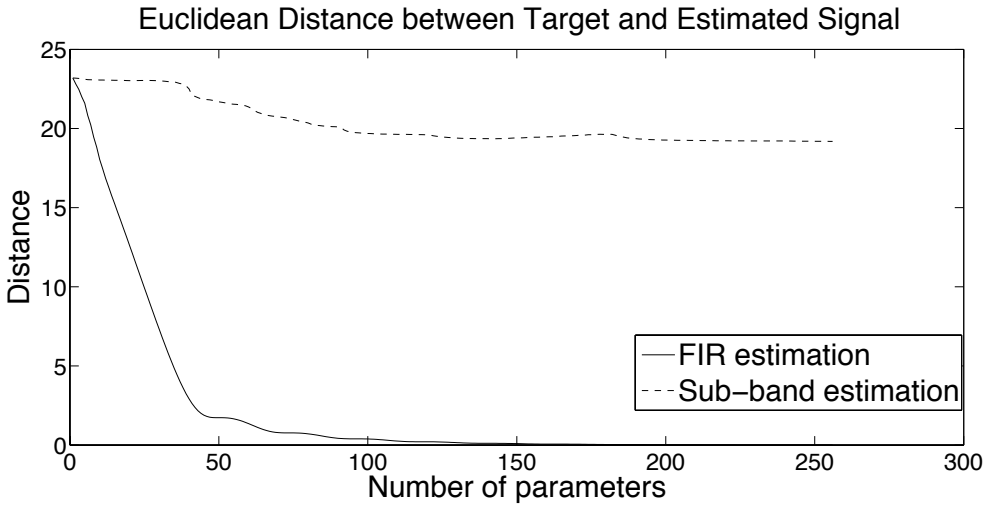

Figure 8: *Distance from target for an 8th order IIR lowpass filter*

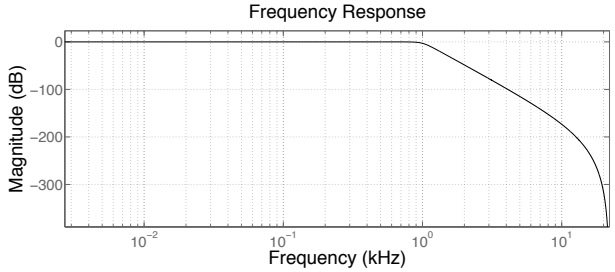

Figure 9: *Frequency response of an 8th order IIR lowpass filter*

Figure 10 shows the computational time required for the estimation of parameters for a 1024-samples audio signal. The length of the test signal for this experiment has been chosen to be one of the typical values used in a frame-by-frame implementation.

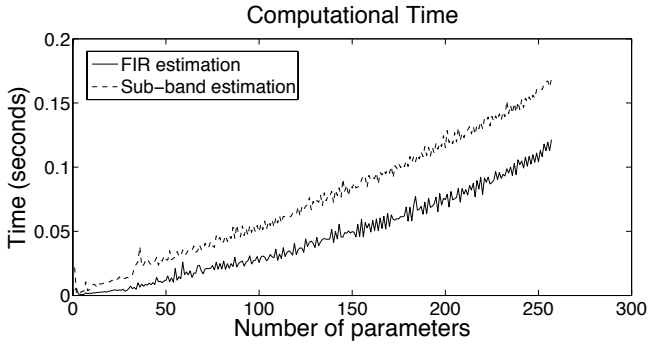

Figure 10: *Computational time required for parameters' estimation*

The FIR estimator works completely in the time domain, while the sub-band method requires the computation of the Fourier transforms of each channel of the multi-track recording. Therefore the computational time required by the FIR technique is always smaller.

In order to demonstrate the effectiveness of the FIR estimator with a multi-track recording, six different 128-order FIR filters have been applied to six different channels, and then the channels were mixed to produce the target mix.
The filters have a "free-hand" impulse response which is the result of choosing random gains for 8 particular frequencies and designing a 128-order FIR filter whose response is an interpolation between these points.
Figure 11 shows the response of one of the filters. The frequencies specified by the dashed lines have a fixed gain level between $+12dB$ and $-12dB$.

The FIR estimator is able to retrieve different filter coefficients for each track, and figure 12 shows the results in terms of the distance between target and estimated mix.
Although the distance doesn't drop exactly to zero as the number of parameters reaches the order of the target filter, the error becomes very small and the estimated mix can be considered to be almost equal to the target. This error may be due to the great number of operations involved in the estimation, and may be reduced by performing the algorithm on a frame-by-frame basis.

5. CONCLUSIONS AND FURTHER RESEARCH

A new method for automatic retrieving of mixing parameters has been proposed which improves the precision, the

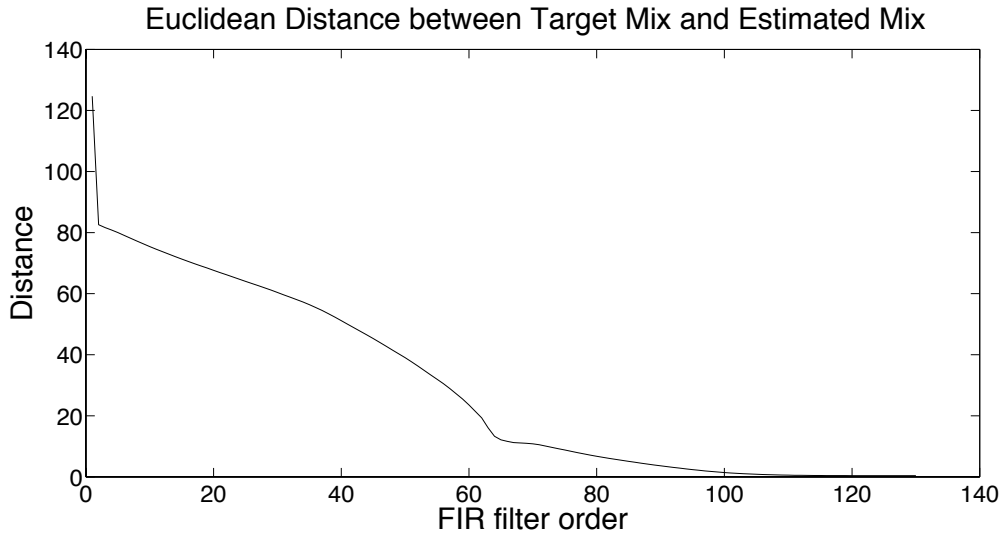

Figure 12: *Distance from target for a 6-track recording, equalized with one different free-hand FIR filter for each track*

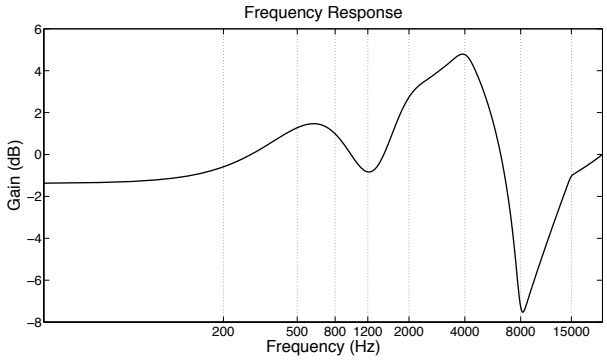

Figure 11: *Frequency response of a 128th order Free-hand FIR filter*

robustness and the computational cost of previous work [6] and takes into account equalization parameters.

As discussed in section 1, a possible direction for further research aims at solving the general problem when target and multi-track recording contain different audio tracks which would result in the automation of the mixing of novel recordings. In this case our system is not directly applicable because the estimated parameters would depend not only on the actual mixing parameters, but also on the content of target and multi-track recordings.
Having said that, it is possible to perform an optimization of parameters defining different objective functions, which may take into account perceptual similarity between target and estimated mix. In this case the system will build a mix which is percieved as similar to the target, even if the parameters applied are different from those used to construct it.

Our current research is focused on the extension of the ATM algorithm to the estimation of dynamic processor parameters in a multi-track mix. For this purpose we are applying the algorithm on a frame-by-frame basis, assuming that the parameters are constant within a small analysis window. This frame based implementation may also be used to retrieve time varying gains and equalization parameters. However preliminary results indicate that this extension is not straightforward.

6. REFERENCES

[1] Bob Katz, *Mastering Audio: The Art and the Science*, Number ISBN 0240805453. Focal Press, 2002.

[2] Nine Inch Nails, "Remix," http://remix.nin.com.

[3] Stephen Travis Pope and Alex Kouznetsov, "Expert mastering assistant," Tech. Rep., FastLab, Sep 2008.

[4] TC Electronics, http://www.tcelectronic.com/, *Assimilator Manual*.

[5] Dale Reed, "A perceptual assistant to do sound equalization," in *Intelligent User Interfaces 5th Conference*, Jan 2000.

[6] Bennett A Kolasinski, "A framework for automatic mixing using timbral similarity measures and genetic optimization," in *AES 124th Convention*, 2008.

[7] E Pampalk, S Dixon, and G Widmer, "On the evaluation of perceptual similarity measures for music," in *DAFx 03*, Jan 2003.

[8] David Goldberg, *Genetic Algorithm in Search, Optimization and Machine Learning*, Number ISBN 0201157675. Addison-Wesley, 1989.

[9] Gilbert Strang, *Introduction to Linear Algebra*, Number ISBN 0961408898. SIAM, 3 edition, 2003.

[10] O Kirkeby and PA Nelson, "Digital filter design for inversion problems in sound reproduction," *Journal of the Audio Engineering Society*, vol. 47, no. 7/8, pp. 583–595, July/August 1999.

[11] Richard Greenfield and Malcom Omar Hawksford, "Efficient filter design for loudspeaker equalization," *Journal of the Audio Engineering Society*, vol. 39, no. 10, pp. 739–751, October 1991.

[12] Richard Lee, "Simple arbitrary iirs," in *AES 125th Convention*, 2008.