ON THE USE OF MASKING FILTERS IN SOUND SOURCE SEPARATION

Derry FitzGerald, Rajesh Jaiswal, *

Audio Research Group, School of Electrical Engineering Systems
Dublin Insitute of Technology
Dublin, Ireland
derry.fitzgerald@dit.ie

ABSTRACT

Many sound source separation algorithms, such as NMF and related approaches, disregard phase information and operate only on magnitude or power spectrograms. In this context, generalised Wiener filters have been widely used to generate masks which are applied to the original complex-valued spectrogram before inversion to the time domain, as these masks have been shown to give good results. However, these masks may not be optimal from a perceptual point of view. To this end, we propose new families of masks and compare their performance to generalised Wiener filter masks using three different factorisation-based separation algorithms. Further, to-date no analysis of how the performance of masking varies with the number of iterations performed when estimating the separated sources. We perform such an analysis and show that when using these masks, running to convergence may not be required in order to obtain good separation performance.

1. INTRODUCTION

In recent years, there has been much work on sound source separation (SSS) algorithms which operate on magnitude or power spectrograms, including those approaches based on Non-negative matrix factorisation and other related approaches [1, 2]. A problem with such approaches is that there are no phase information available for the separated source spectrograms to allow inversion to the time-domain. To overcome this problem, a number of different approaches have been employed. Phase estimation techniques, such as that of Griffin and Lim [3] , or more recently that of Le Roux et al [4] have been used. An alternative approach was to simply reuse the phase of the original mixture signal when resynthesising. However, the most commonly used approach in recent years has been to use the estimated source spectrograms to create generalised Wiener filters, which are then used as soft masks to be applied to the original complex valued spectrogram. This approach can be formalised as:

$$\mathbf{X}_k = \mathbf{Y} \otimes \mathbf{M}_k \tag{1}$$

where for generalised Wiener filters, the soft mask $\mathbf{M}_k$ is defined as:

$$\mathbf{M}_k = \frac{\mathbf{S}_k^r}{\sum_{p=1}^{P} \mathbf{S}_p^r} \tag{2}$$

Here $\mathbf{X}_k$ is the estimated complex spectrogram of the kth source, $\mathbf{Y}$ is the original complex mixture spectrogram, $\mathbf{S}_k$ is the estimated spectrogram of the kth source,and P is the total number of sources. The exponent r is 1 for power spectrograms, or 2 for magnitude spectrograms. All divisions are elementwise throughout the remainder of the paper and $\otimes$ denotes elementwise multiplication.

The generalised Wiener filter was initially proposed by Benaroya et al [5] in the context of single channel separation. Since then it has been used in numerous sound source separation algorithms including [6], where it was used in the context of drum sound separation, user-assisted separation in [7], and for source-filter based separation in [8].

In effect, this approach allocates the energy in a given time-frequency bin across the sources according to a least-squares best fit. Another advantage of this approach is that the separated sources sum together to give the original mixture signal. This is of particular benefit for remixing purposes, or for upmixing from mono to stereo for example. Here the artefacts and errors in separation will often be masked due to the presence of the other sources [9].

To-date little or no attention has been paid to improving the performance of the masking approach, with the notable exception of work by Le Roux et al [10], where they impose spectrogram consistency constraints to obtain better performing masks, leading to improved separation results. It is also worth noting that while the masks may be optimal in the least squares sense, there is no guarantee that the masks generated are optimal from a perceptual point of view, and it may be that other masks are more optimal from a perceptual perspective. Further, no investigation has been made on how the performance of the masks vary with the number of iterations performed by the separation algorithm. Instead, the masks have only been applied on completion of the algorithm. It is proposed to investigate these issues in the remainder of this paper.

Section 2 proposes a new set of masks for use with SSS, with section 3 then outlining the testsets and algorithms used for testing these new masks. Section 4 then contains results and discussion on the tests performed using these new masks. Finally, section 5 highlights conclusions drawn and areas for future work.

2. DIVERGENCE-BASED MASKS

As noted previously, generalised Wiener filtering is optimal in a least-squares sense. However, in the context of sound source separation algorithms, particularly those based on NMF, least-squares approximations have typically been outperformed by other cost functions. In particular, two widely used cost functions are the Kullback-Leibler divergence, and the Itakura-Saito divergence. The generalised Kullback-Leibler (KL) divergence is given by:

$$D_{KL}(\mathbf{X}, \mathbf{Y}) = \sum \mathbf{X} \log \frac{\mathbf{X}}{\mathbf{Y}} - \mathbf{X} + \mathbf{Y} \tag{3}$$

* This work was supported by Science Foundation Ireland

where summation takes place over all elements of $\mathbf{X}$ and $\mathbf{Y}$. The Itakura-Saito (IS) divergence is given by:

$$D_{IS}(\mathbf{X}, \mathbf{Y}) = \sum \frac{\mathbf{X}}{\mathbf{Y}} - \log \frac{\mathbf{X}}{\mathbf{Y}} - 1 \tag{4}$$

These cost functions have been found to perform well when used with NMF-based approaches to separate sounds sources, giving better results than using a least-squares based cost function. Therefore, we propose to develop masks based on these divergences to see if they outperform the generalised Wiener filter mask.

To this end, we define a family of divergence-based masks:

$$\mathbf{M}_k = 1 - \frac{D(\mathbf{S}_k, \mathbf{Q})^t}{\sum_{p=1}^{P} D(\mathbf{S}_p, \mathbf{Q})^t} \tag{5}$$

where $\mathbf{M}_k$ is the mask associated with the kth source, $\mathbf{Q}$ is the estimated mixture spectrogram, and exponent t is used to vary the properties of the mask. Here D denotes any suitable divergence metric. It should be noted that both the Kullback-Leibler divergence and Itakura-Saito divergence tend towards zero when the datapoints are similar, and so the term after the minus sign in eqn. (5) defines a mask which removes the source from the mixture. Subtracting this from a value of 1 then yields the mask to separate the source in question. The complex source spectrograms are then estimated as per eqn. (1), but with the chosen mask instead of the generalised Wiener filter.

Like the generalised Wiener filter, sources separated using these masks will sum together to reconstruct the original mixture signal, making them suitable for remixing or upmixing purposes. However, here the energy in a given bin is now allocated based on goodness of fit to the chosen divergence metric.

3. MASKING TESTSETS

We propose to evaluate the performance of the divergence-based masks for both the KL and IS divergences, and for t values of 1 and 2. We use three different algorithms with their associated testsets to evaluate the performance. This is in an attempt to ensure that the results obtained are not specific to a given algorithm.

The first algorithm used is the Source-Filter Sinusoidal Shifted Non-negative Tensor Factorisation (SFSSNTF) algorithm as described in [11]. This algorithm is additive-synthesis based and assumes that a given instrument can be modelled as a frequency invariant set of harmonic weights, in conjunction with a formant filter which allows the timbre of the instrument to change with pitch. The testset used consisted of 25 mono mixtures of two pitched sources, and this testset is publicly available at [12]. Details on the creation of this testset can be found in [11] .

The second algorithm is the user-assisted source separation algorithm (UA) described in [13]. Here, the user sings along with the source to be separated. This recording is then factorised using NMF, and the resultant basis functions are then used as priors to guide the factorisation of the mixture signal in order to extract the desired source. The testset used here was created from a set of recordings by the Beach Boys where the vocals and accompanying backing track were available separately [14]. These were manually synchronised and then mixed down to a mono mixture. User guides were then recorded of a user singing along with the vocal melody. Further details on this testset can be found in [15] and the influence of the priors was gradually removed over the course of the first 20 iterations, so that after 20 iterations the update equations collapsed to those of standard NMF. 100 basis functions were used to capture the target source, i.e. the vocal, while a further 100 were used to model the backing track.

The third algorithm is NMF followed by a clustering stage based on Shifted NMF which clusters the NMF basis functions to sources, as described in [16]. The testset used is the same as that for the SFSSNTF algorithm. Here, NMF followed by SNMF was used to identify the clustering, and then NMF was ran again, using the same initialisation as the original NMF stage, so that the second NMF will converge to the same point as the original NMF, and so the clustering obtained via SNMF still applies. As no constraints have been applied to NMF, this will demonstrate the effects of masking when used with a standard NMF algorithm.

It can be seen that, while all based on NMF, the algorithms achieve separation using very different methods and constraints. Similarly, the test sets are very different in nature, and so the results obtained regarding the masks should generalise well.

The three algorithms were ran for 100 iterations using magnitude spectrograms and the KL divergence as a cost function. All test signals were mono mixtures with a sample rate of 44.1 kHz, and an FFT/window size of 4096 samples and a hopsize of 1024 samples were used. The separation performance of the masks was evaluated after every 10 iterations. Also evaluated as a baseline was the generalised Wiener filter mask.

Evaluation of the performance of the masks was done using the PEASS toolbox, which calculates a set of objective measures for the perceptual evaluation of audio source separation [17]. The metrics used were the overall perceptual score (OPS), the target-related perceptual score (TPS), the artifacts-related perceptual score, and the interference-related perceptual score (IPS).

OPS measures the perceived overall quality of the separation, while TPS measures how well the separated source matches the spatial positioning of the original source. IPS determines how much interference due to other sources is perceived in the separated source, and finally APS measures the perceived amount of artifacts in the separated source.

4. EVALUATION

Figure 1 shows the average OPS values obtained for SFSSNTF and its associated testset. A circle-dashed line indicates the generalised Wiener filter mask, stars indicates the use of the KL divergence mask, with diamonds indicating the IS divergence mask. A solid line indicates the use of $t = 2$ for the associated divergence mask, while a dotted line indicates $t = 1$. The same holds for all subsequent figures.

It can be seen that the proposed masks all outperform the generalised Wiener filter, with the KL mask with $t = 2$ performing best. It is very interesting to note that the best performance is obtained at 10 iterations, long before the algorithm has converged, and that the general trend after is a decline is OPS.

Figure 2 shows the average OPS results for UA on its testset. Again it can be seen that the proposed masks outperform the generalised Wiener filter, but that now the KL mask with $t = 1$ performs best, followed by the KL mask with $t = 2$. It is also interesting to note that the best performance in terms of the metrics for all masks is obtained at 20 iterations, which is the point where the influence of the guide priors is been removed from the update

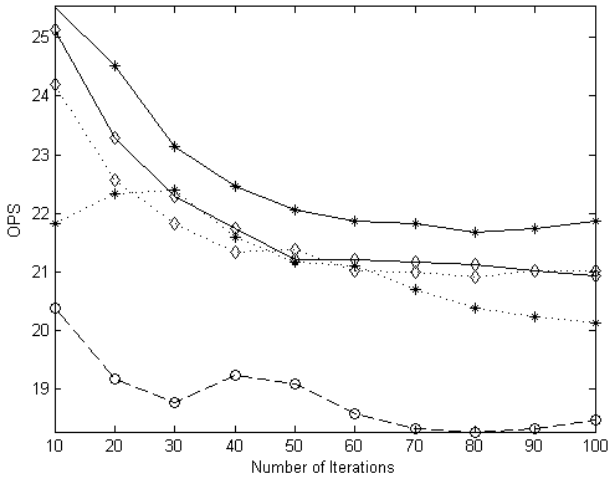

Figure 1: *Overall Perceptual Scores for the Source-Filter Sinusoidal Shifted Non-negative Tensor Factorisation algorithm. Circle dashed indicates the use of the generalised Wiener filter, stars indicate the use of a KL divergence based mask, while diamonds indicate the use of an IS divergence based mask. A dotted line indicates p = 1 while a solid line indicates p = 2 when generating the masks. The same legend is used for all subsequent figures.*

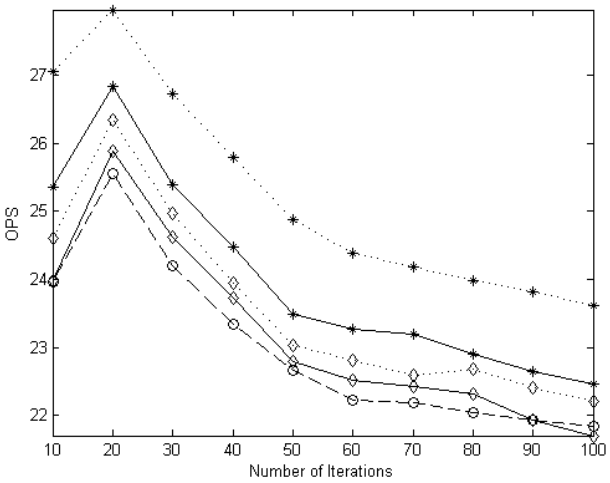

Figure 2: *Overall Perceptual Scores for the User Assisted algorithm, Legend as per Figure 1.*

equations. This suggests that after this point the basis functions begin to adapt to capturing extra details in the overall mixture, rather than optimising the individual sources to be separated.

Figure 3 shows the average OPS results for standard NMF on its testset. As with the previous algorithms, it can be seen that the OPS scores peak long before numerical convergence is achieved, at a value of 50 iterations . At this value, it can be seen that the proposed new masks again outperform the generalised Wiener filter, though with a smaller improvement in performance than with the previous two algorithms. Here the masks with p=2 perform best, with the IS mask slightly outperforming the KL mask. This suggests that the KL mask with $t = 2$ is the mask which generalises best in terms of OPS as it performs consistently well for all algorithms, being either first or second in performance.

Figures 4,5 and 6 show the TPS scores for the SFSSNTF, UA and NMF algorithms respectively. Similar trends can be observed for TPS as for OPS for the SFSSNTF and UA algorithms , though the best performing masks are different, with the generalised Wiener Filter performing well. For the standard NMF algorithm TPS can be seen to increase gradually with increasing iterations. The KL mask with $t = 2$ here comes in the middle in terms of performance.

Figures 7,8 and 9 then show the results obtained for IPS for SFSSNTF and UA respectively. SFSSNTF shows a downward trend for IPS, while most of the metrics again peak at 20 iterations for UA. For standard NMF, the values vary with iteration number. Here, the KL mask with $t = 2$ is again in the top two masks, with the IS mask with $t = 2$ performing best overall, suggesting that if rejection of the other source is the priority, then this is the optimal mask to use.

Finally, figures 10, 11 and 12 show the APS scores obtained. Here, the KL mask with $t = 1$ is the best performing, followed by

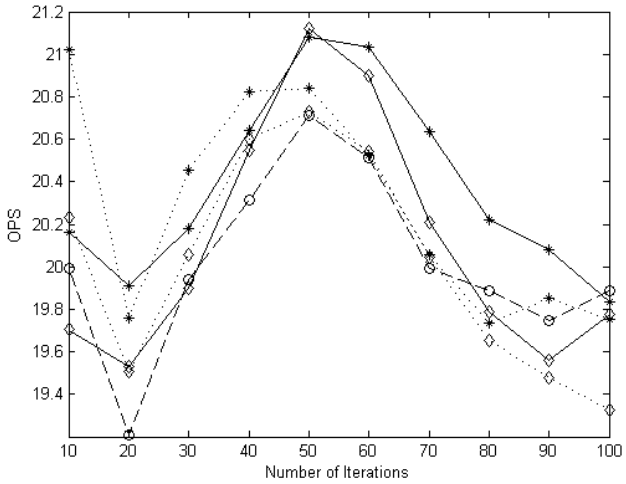

Figure 3: *Overall Perceptual Scores for the standard NMF algorithm, Legend as per Figure 1.*

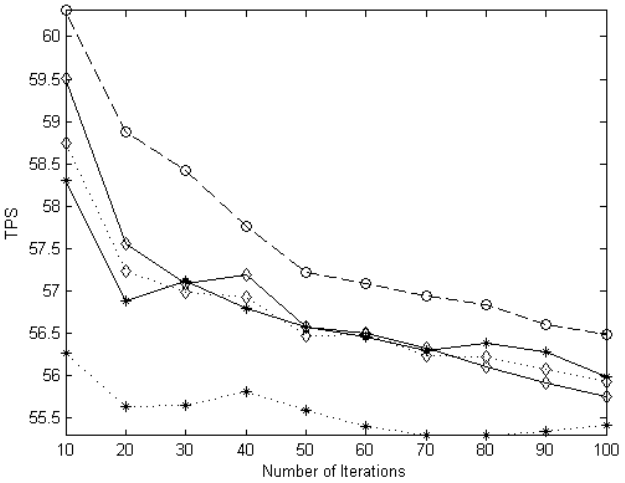

Figure 4: *Target-related Perceptual Scores for the Source-Filter Sinusoidal Shifted Non-negative Tensor Factorisation algorithm. Legend as per Figure 1.*

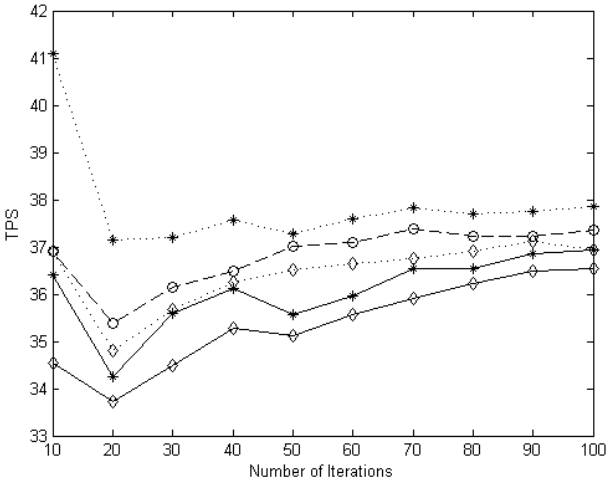

Figure 6: *Target-related Perceptual Scores for the standard NMF algorithm. Legend as per Figure 1.*

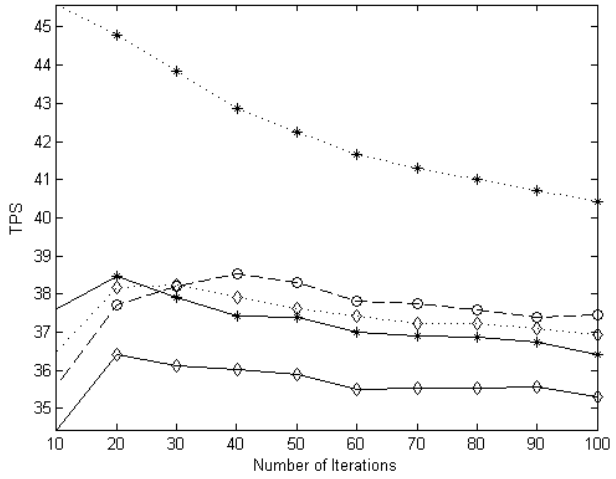

Figure 5: *Target-related Perceptual Scores for the User Assisted algorithm. Legend as per Figure 1.*

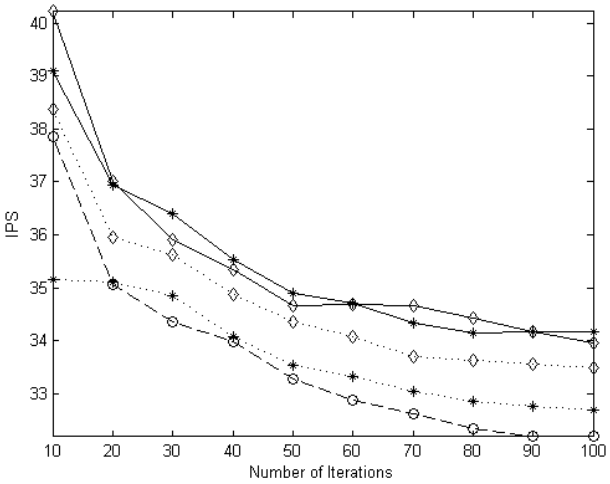

Figure 7: *Interference-related Perceptual Scores for the Source-Filter Sinusoidal Shifted Non-negative Tensor Factorisation algorithm. Legend as per Figure 1.*

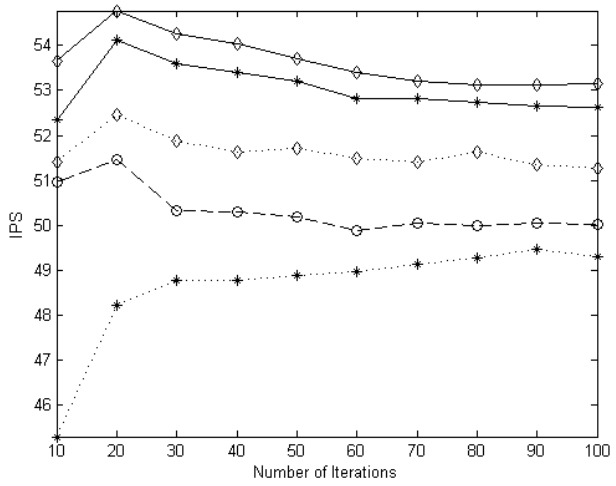

Figure 8: *Interference-related Perceptual Scores for the User Assisted algorithm. Legend as per Figure 1.*

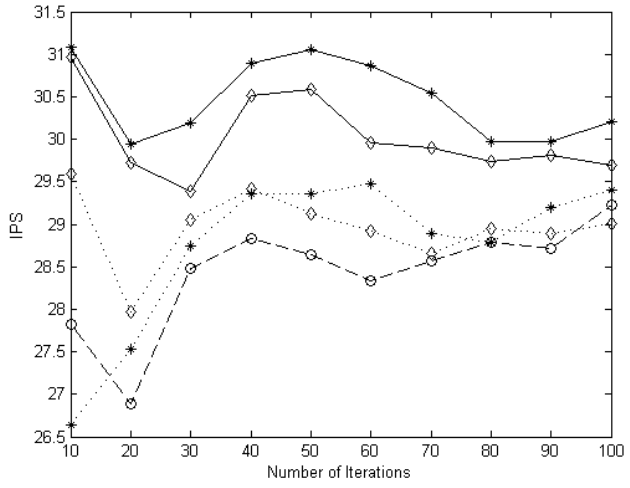

Figure 9: *Interference-related Perceptual Scores for the standard NMF algorithm. Legend as per Figure 1.*

the generalised Wiener filter. It can be seen that both the KL and IS masks with $t = 2$ perform worst here, suggesting that there is a trade off between source rejection and the presence of artifacts in the separations, with more rejection resulting in increased artifacts.

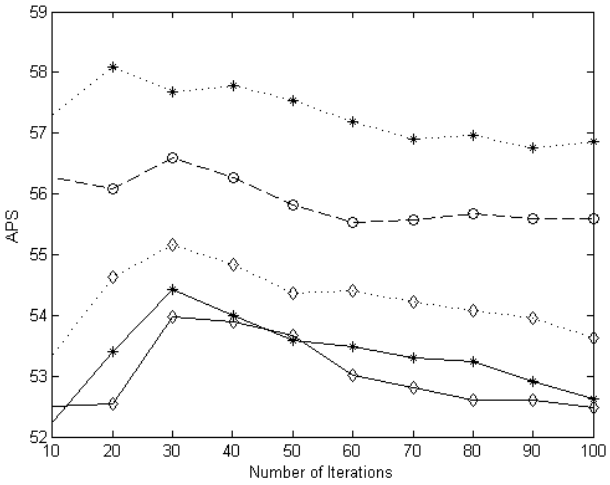

Figure 10: *Artifacts-related Perceptual Scores for the Source-Filter Sinusoidal Shifted Non-negative Tensor Factorisation algorithm. Legend as per Figure 1.*

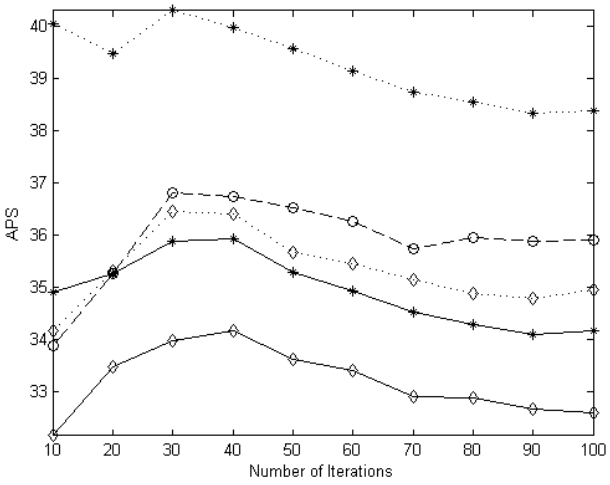

Figure 11: *Artifacts-related Perceptual Scores for the User-Assisted algorithm. Legend as per Figure 1.*

Overall, it can be seen that no individual mask performs equally well across all the metrics, suggesting that the mask should be chosen according to the purpose for which the separation is required. If the best overall separation is required then the KL mask with $t = 2$ is most likely to give best results, while if suppression of the other sources is required, then the IS mask with $t = 2$ is best. Nonetheless, in terms of OPS, the masks proposed all outperform the widely used generalised Wiener filter.

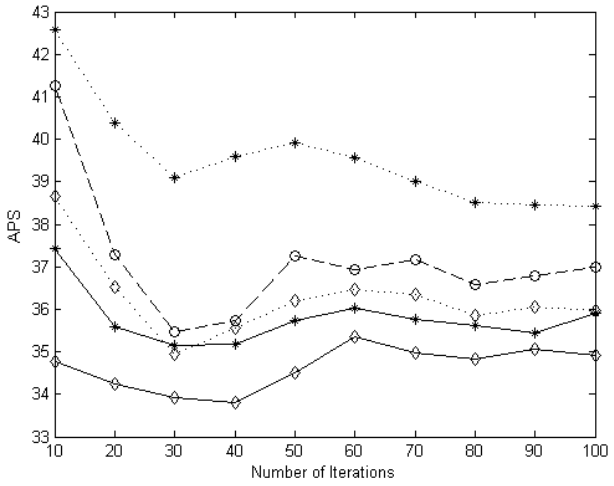

Figure 12: *Artifacts-related Perceptual Scores for the User-Assisted algorithm. Legend as per Figure 1.*

An unexpected and surprising result from these tests is the fact that the OPS scores were highest at low numbers of iterations, long before the algorithms have converged. This is contrary to what would be expected, which is that more accurate modelling of the sources would lead to more accurate separation. Further investigation revealed that this is in fact the case, provided that the sources are resynthesised from the estimated spectrograms directly (here we used the original mixture phase), rather than used to generate a mask to apply to the mixture spectrogram. Therefore it follows that the high OPS scores at low iterations are as a direct result of the use of masking when reconstructing the signals.

A potential explanation for this can be found in the fact that audio spectrograms are sparse in nature. Therefore, there will be many bins where there is little or no energy present. In contrast, due to the random initialisation of the basis functions, the corresponding bins in the estimated source spectrograms are initially likely to contain significant energy. This is because, at low numbers of iterations, the basis functions will not have adapted enough to remove this energy. However, if these estimated source spectrograms are used to generate masks then the energy in these bins no longer matters. The masks allocate energy in the original spectrogram in proportion to that of the source estimates, and a proportion of a small number only yields a smaller number. Therefore, the masks can be seen to eliminate noise present in the estimated source spectrograms obtained at low numbers of iterations, particularly for bins with low energy in the original mixture.

For bins with significant energy the above reasoning does not hold. However, these bins will be further away in magnitude from the initial values obtained from the random initialisation, and so the rescaled gradient in the multiplicative updates used in these algorithms will be larger, with the result that these bins are more likely to converge faster, at least over the initial iterations. Therefore, at low numbers of iterations, these bins are more likely to contain reasonable estimates of the actual source energy than the other bins. Further, with the use of masking, what is important is the proportion of energy, not the actual energy present. Once the proportion is approximately correct, then good separation at a given bin can be obtained regardless of errors in the actual energy estimates or the number of iterations performed. It can then be seen that obtaining good separation at low iteration numbers, while initially seeming counter-intuitive, makes sense once the effects of masking have been taken into account. Therefore, it may not be necessary to run these factorisation based algorithms to convergence in order to obtain the best separation performance. This offers the potential to greatly reduce run-times for these separation algorithms while still obtaining good separation performance. Audio examples demonstrating the effects of masking on separation performance can be found at [18].

5. CONCLUSIONS

Having discussed the use of the generalised Wiener filter as a means of resynthesis when performing SSS, we then noted that while optimal in a least-squares sense, there is no guarantee that the mask is optimal from a perceptual point of view. To this end, a new family of masks based on the Kullback-Leibler and Itakura-Saito divergences were then introduced. These masks were shown to outperform the generalised Wiener filter for overall separation quality in terms of perceptually motivated separation metrics when tested using three different separation algorithms and two datasets. It was also demonstrated that good separation performance can be obtained at low numbers of iterations, suggesting that it is not necessary to run to convergence to get good separation performance with these algorithms. Areas for future work include extending the family of masks to include the Beta divergence to attempt further improvements. Given the effects that masking has on the outputs, it is proposed to investigate incorporating the masks in the optimisation process to see if separation results can be improved.

6. REFERENCES

[1] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Proc. of Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[2] A. T. Cemgil T. Virtanen and S. J. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[3] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE transactions on Acoustics, Speech and Signal Processing*, vol. 32, 1984.

[4] J. LeRoux, N. Ono, and S. Sagayama, "Explicit consistency constraints for stft spectrograms and their application to phase reconstruction," in *Proc. SAPA 2008 Workshop on Statistical and Perceptual Audition (SAPA 2008)*, 2008.

[5] L. Benaroya, L. Mc Donagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[6] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 529–540, 2008.

[7] P. Smaragdis and G. Mysore, "Separation by humming user guided sound extraction from monophonic mixtures," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA09)*, 2009.

[8] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and em algorithm," in *35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[9] D. FitzGerald, "Upmixing from mono - a source separation approach," in *17th International Conference on Digital Signal Processing*, 2011.

[10] J. LeRoux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010)*, 2010.

[11] D. FitzGerald, M. Cranich, and E. Coyle, "Extended non-negative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience, vol. 2008, Article ID 872425*, 2008.

[12] D. FitzGerald, "Test database," Webpage, `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=52`.

[13] D. FitzGerald, "User assisted source separation using non-negative matrix factorisation," in *IET Irish Signals and Systems Conference*, Dublin, 2011.

[14] "The pet sounds sessions," The Beach Boys, 1997, Capitol 72438 37662 2 2.

[15] D. FitzGerald, "Nmf-based algorithms for user assisted sound source separation," *submitted to EURASIP Journal on Audio, Speech, and Music Processing*, 2012.

[16] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and S. Rickard, "Clustering nmf basis functions using shifted nmf for monaural sound source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[17] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[18] D. FitzGerald, "On the use of masking filters in sound source separation," Webpage, `http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489&menu_id=55`.