2017 UKSim-AMSS 19th International Conference on Modelling & Simulation

Mask Optimisation for Neural Network Monaural Source Separation

Richard Cant, Caroline Langensiepen, William Metcalf
School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom
Richard.Cant@ntu.ac.uk

***Abstract*—An ideal binary mask is a means by which multiple sound sources within a single audio file can be separated. Previous work has shown a deep neural network can be trained to approximate the ideal mask, but at a substantial computational cost. We present a method to assess the impact of reducing the mask by averaging time and frequency bins, so that the computational cost can be significantly reduced. Our work uses the original separate musical channels mask as a ground truth and compares this against an ideal binary mask and an ideal ”soft” or proportional mask. The ideal soft mask is then compared against masks produced by a range of averaging levels. We find that averaging could produce a reduction by a factor of 16 in the number of weights in the neural network (and thus a significant improvement in computation time), while still achieving plausible results in terms of source separation.**

Keywords–signal processing; optimisation; neural networks

I. Introduction

Multitrack audio recordings have provided studio engineers with fine control over post production for many years. With many studios now digital (for cost and convenience reasons), a large proportion of modern pop music is recorded with tens of individual sources. However, many of the outputs are exported as a single track, the original sources remaining under lock and key in the mixing engineers editing booth. This is particularly unfortunate, as an increasing number of consumers now have access to affordable advanced audio editing software and hardware, with the potential to take advantage of these digital recordings in novel ways.

One of the most common questions asked by the consumers is how a vocal portion of a mix can be isolated from the instrumental. Consumers desire to do this for two primary reasons: either to use the instrumental as a karaoke style backing, or more commonly, to create ‘remixes’ of their own by overlaying vocals on new instrumental beds. This is an example of signal source separation: isolating and recovering specific signals which have been mixed into a single combined source. The ultimate aim is to recover the original component without interference or noise from other combined sources. In audio signal processing, this is an example of the ‘cocktail party’ problem [1], [2] whereby multiple voices are captured simultaneously. The human ear is remarkably good at isolating these voices, using gender, spatial information, amplitude and other factors for separation. Attempting to reproduce this using digital signal processing is much more difficult, with multiple approaches tackling different variants of the problem.

This paper explores means by which a recent strategy for addressing this problem could be optimised to achieve the goals within more reasonable computational costs. Section II discusses various techniques that have been applied to the cocktail party. Section III explains the motivation behind the current work, and how it uses the data from the original method to work out the possible impact of optimisation on the outcomes of the neural network. Section IV presents the results, the way in which they are assessed and analyses their quality, while suggestions for further exploration are covered in section V.

II. Previous Work

Separation of multiple sources within an audio signal has been attempted by a variety of means. For example, Non-negative matrix factorisation (NMF) considers the original combined signal as a matrix, and tries to find two smaller matrices that can be multiplied together to give a good representation of the original signal [3]. However this has problems where the signals do not form well-separated clusters (i.e. have some overlap). Grais et al [4] and Kang et al [5] advanced on this by using different deep neural network approaches in conjunction with NMF.

A somewhat different strategy is the development of the ‘ideal binary mask’ approach. This was first suggested in 2001 by Hu and Wang [6]. It has been argued that the ideal binary mask approaches the SNR performance of the ideal ratio mask, and thus is close to the theoretically optimal Weiner filter [7]. The ideal binary mask has since been explored extensively, most noticeably in the area of improving the intelligibility of speech within a noisy environment [8], [9]. Simpson et al [10] used the ideal binary mask to train a deep neural network on the cocktail party problem. They found that the deep neural network is capable of separating out vocal parts from a musical mixture, and that it has the advantage of being able to learn what ‘vocal’ sounds consist of.

978-1-5386-2735-8/17 $31.00 © 2017 IEEE
DOI 10.1109/UKSim.2017.21

III. Experimental Method

Simpson et al [10] used a neural network of 3 layers with a size of 1025×20 each. This produces some 840 million nodes and over 1 billion parameters - which constituted a significant computational load for training: in private communication with the authors, they indicated that “training takes about a month on a big single core server”. [11] Several studies have highlighted calculation time as a limiting factor, often choosing to process only a small subset of test data as a result. It is clear that this would be a major impediment to any real-world application [12]. Our aim is therefore to assess the ways in which the computation time could be shortened by reducing the number of input and output nodes in the neural network. The object of the present work is to examine the effect of this reduction on the resulting sound quality.

Our starting hypothesis is that a perfectly performing neural network would, after training, produce an output that is indistinguishable from the ideal binary mask. We therefore consider the ideal binary mask (that would be used to train the neural network) as the best possible output that could be achieved by applying the ‘perfectly trained’ network to a new piece of music. We generate this ideal binary mask from the composite, instrumental and voice tracks as was done in ‘Deep Karaoke’. We then generate our optimised masks from the same tracks directly. This produces the equivalent of a fully trained neural network working with the reduced input, hidden and output nodes. We then reconstruct the separate instrumental and vocal tracks by applying the masks to the original composite track. This produces 2 sets of separated tracks - one from the ideal mask, and one from our optimised mask. We then compare the outputs from these two processes to assess the quality of the masks. We report both objective measures (e.g. signal to noise ratios) and subjective assessment. The results of these experiments thus produce an upper bound on the quality that could be achieved by the reduced size neural network, in that the optimised maps thus constructed are what a perfectly trained neural network with a reduced set of nodes could achieve at best. This experimental process is illustrated in Figure 1.

As in Simpson [10] the generation of the ideal binary mask was achieved by windowing and Fourier transformation to produce a spectrogram containing 20 time-windowed samples, each containing 1025 frequency components. The magnitudes of the source (vocal, instrumental) parts were then compared with the corresponding magnitudes in the mix spectrograms, generating a 1 where the vocal magnitude was greater, or 0 if the backing magnitude was greater, thus producing an ideal binary mask of 20×1025 elements. Phase information was not used in the mask generation process.

For the generation of the optimised masks, a similar process was applied. To reduce the number of nodes in any neural network, it would be necessary to reduce the frequency points and/or the time samples. Our experiments covered reducing the original 1025 frequencies to 512, 341, 256, 205, 170 and 128. The number of time samples was also reduced from 20 to 10, 5 and 2. In each case the input data was calculated by simple averaging (over frequency groups or time intervals) and the output (mask) value was used repetitively across multiple samples. The Fourier transforms that constructed the samples and reconstructed the audio afterwards always used the full resolution and the phase information.

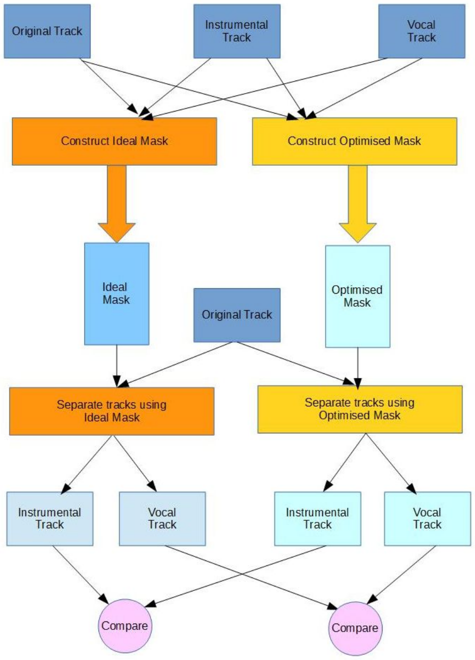

Figure 1. Experimental process of constructing masks for comparison

IV. Results and Analysis

The results of this type of exercise can be assessed in two ways. The simplest approach is to use objective criteria based on signal to noise ratio. However it is well known in this field that such an approach is not adequate. In fact the relationship between subjective and objective assessment has formed a substantial field of study in its own right, for example [13]. Unfortunately to perform a full subjective assessment is a substantial undertaking so, for simplicity, we will present objective results based on signal to noise ratio supported by our own subjective assessment.

To generate our results, 7 musical mixtures were first chosen from the Medley DB library [14]. They were selected for their difference in genre and style, including pop, rock and indie within the test set. This was to evaluate the effectiveness across audio with differing number of stems and styles. Each track was reduced to 60 seconds in length and converted to mono for the purpose of simplicity; reduced in gain by -3db to avoid overshoots (importantly not normalised

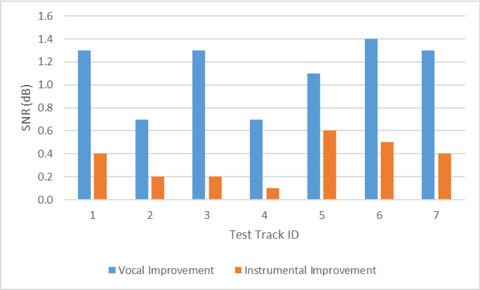

Figure 2. Distinct SNR improvement achieved by binary mask over using using a proportional or ”soft” mask, both at full 20 × 1025

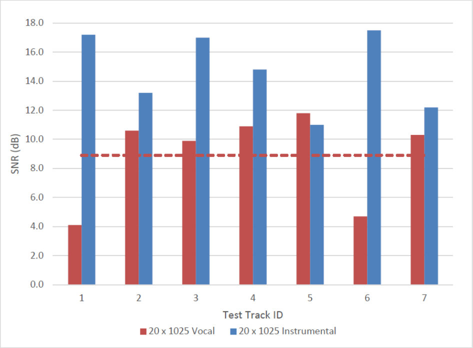

Figure 3. Signal to noise ratio of reconstructed data at full 20 × 1025 compared to original

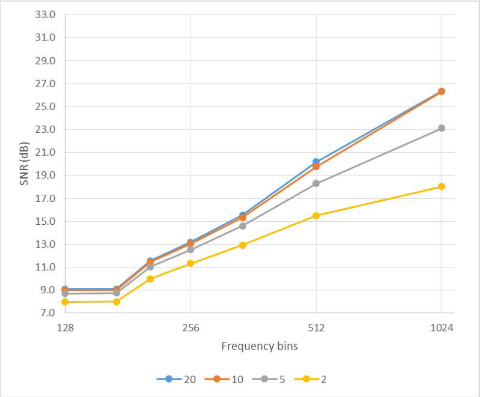

Figure 4. Mean vocal SNR for all musical mixtures, samples (time) vs frequency bins

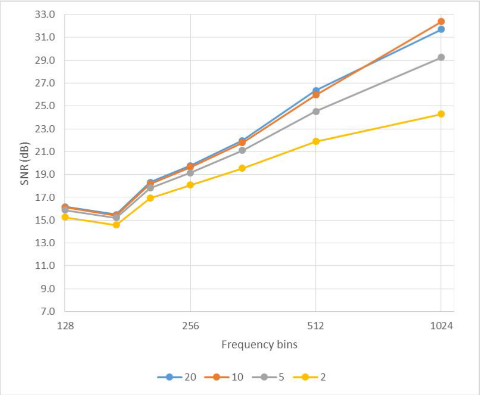

Figure 5. Mean instrumental SNR for all musical mixtures, samples (time) vs frequency bins

to retain original stem balance); and then mixed down to produce 3 individual parts, from now on referred to as “vocal”, “instrumental” and “mix”. Each part was exported as PCM, uncompressed WAV with 16 bit depth at a 44100Hz sample rate. This produced 3 parts for each of the 7 test musical mixtures, which were passed through the application to generate various levels of averaging. Generated files were then imported into the Audacity application. The original vocal and instrumental tracks were inverted and mixed with the equivalent vocal or instrumental generated from the full 20 × 1025 mask. The ratio of the RMS of this mix with the original represents the best achievable SNR assuming a perfectly generated mask at the full resolution.

In [10] the neural network output was a “soft” mask with vocal and instrumental contributions for each time-frequency bin being real numbers in the range 0-1. However they introduced a parameter α that acted as a cut off in order to construct a binary mask. Initially we tried using a binary mask with a value of $\alpha = 0.5$ and compared the result with direct use of the original soft mask. The relative signal to noise ratio (SNR) of the two different approaches, each compared to the original ground truth data, is shown in Figure 2. It can be seen that the binary mask shows a considerable improvement over the soft mask. However a subjective assessment of the sound quality produced exactly the opposite result. The binary mask produced an “underwater” sound that was audibly quite unpleasant. The authors of [10] performed an optimization procedure to determine the value of α. We have to assume that this procedure would have eliminated this effect. For simplicity we decided to generate the remainder of our results for the soft mask only, partly because of the better subjective quality, but also to avoid the need to re-optimize α for every combination.

To isolate the effects of the averaging procedure we compare the averaged results, not with the original ground truth, but with the data generated at the maximum of 20×1025. The reference against which the rest of our results are compared is thus the 20 × 1025 soft mask output. The SNR of this output against the original ground truth is shown in Figure 3.

Using this reference, the same process was then repeated for multiple levels of averaging. The resulting SNR represents the degradation that results from the averaging process.

Results are reported as SNR relative to 0dB full scale. A SNR greater than 20db is desired when the separated audio is to be listened to as a standalone. However a lesser SNR may be acceptable when the other part is to be re-introduced

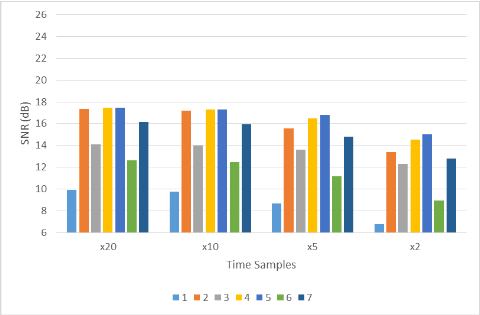

Figure 6. Averaging time samples impacting the vocal SNR for each musical mixture (1 - 7)

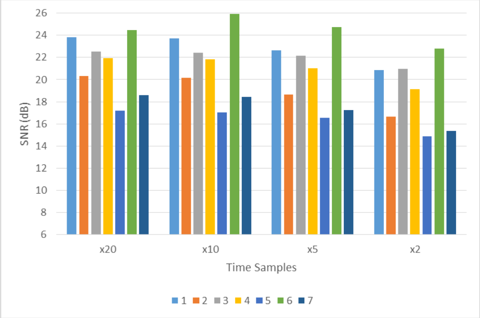

Figure 7. Averaging time samples impacting the instrumental SNR for each musical mixture (1 - 7)

in a new mix.

Figures 4 to 11 show our results for signal to noise ratio for different levels of averaging. The results shown in Figure 4 and Figure 5 summarise the performance of all the test musical mixtures combined, and give a good indication of the average SNR that can be achieved by changing our control parameters (i.e. the number of time samples and frequency bins). They show that reducing the number of time samples from 20 to 10 has very little impact on the SNR achieved, and even at worst (i.e. 20× vs 2×) there was just 4dB of range in the generated vocal samples. In the case of the instrumental samples (Figure 7) the SNR average is much higher, and in the case of test track 6 the SNR only falls from 24.4 to 22.8 even when the number of time samples is reduced from 20 to 2.

The effect of averaging over frequencies appears to be much stronger. These results show a clear and steady reduction in SNR at each level of frequency bin reduction, reaching a SNR below 10dB in some cases. It can also be seen that there are differences between the vocal and instrumental parts. On average, vocal isolation offers a SNR that is inferior by roughly 6dB to that obtained for the instrumental part. This is to be expected, as instrumental parts (typically formed of many layers) are much more difficult to eliminate than the solo voice. This result was also reinforced by subjective assessment since the human ear is also very good at identifying things that sound out of place, such that even the quietest 'beat' over a voice sample can be easily distinguished. When this difference has been taken into account the trend line for the two is extremely similar. The small rise in SNR visible towards the lower frequency bins of Figure 5 could be down to a more accurate reduction factor (i.e. 1025/128 vs 1025/170).

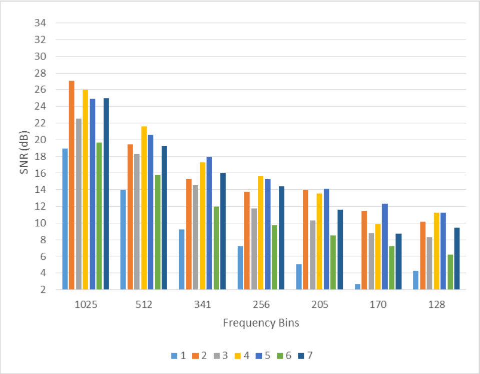

Figure 8. Averaging frequency bins impacting the vocal SNR for each musical mixture (1 - 7)

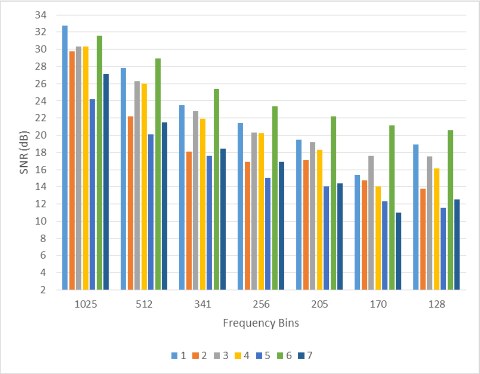

Figure 9. Averaging frequency bins impacting the instrumental SNR for each musical mixture (1- 7).

Figures 6 to 9 show the same data broken down by track. They illustrate the variation in performance that could be expected if the technique were to be applied to different pieces of music in different genres. It can be seen that test track 1 shows particularly poor performance when trying to extract the vocal; however the SNR is not as bad when attempting to remove it and retain only the instrumental.

Therefore, given our target of greater than 20dB SNR, we can now begin to balance the required weights of a typical network with the approximate SNR that could be achieved.

If we assume a typical neural network configuration, where the number of weights are proportionate to our reduction factor, then Figure 4 shows that at 10 time samples and 512 frequency bins, an average of 19.8dB SNR can be achieved. This would reduce the network input / output

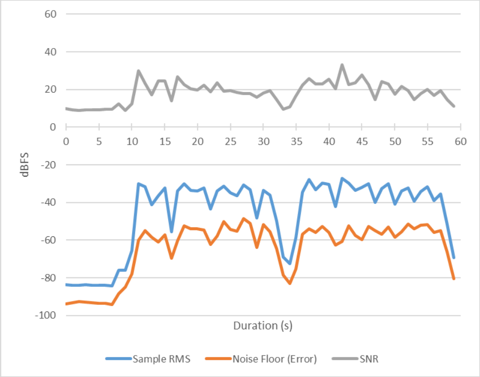

Figure 10. Vocal RMS, noise RMS and resulting SNR for test track 5, at 10 × 512 averaging over the 60 second duration

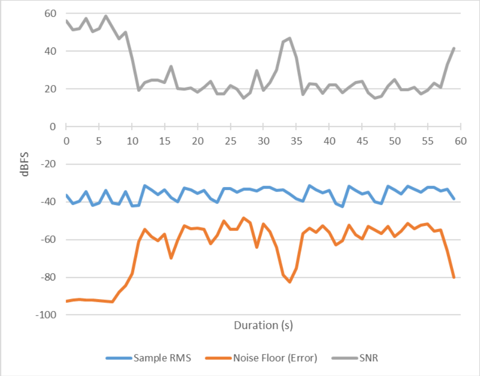

Figure 11. Instrumental RMS, noise RMS and resulting SNR for test track 5, at 10 × 512 averaging over the 60 second duration.

size from 20 × 1025 (20,500) to 10 × 512 (5120). For a simplified network, this would reduce the 840 million (840,500,000) weights to 52 million (52,428,800). In the case of instrumental parts, this can be reduced further. Figure 5 demonstrates that for a filter of 5 × 341, 21.1 dB SNR is achievable. A typical ANN at this level of averaging would require just $2 \times (5 \times 341)^2$ weights, some 5.8 million. Of course, the SNR values used in these figures are averages across 60 seconds of audio, and the SNR at any single point can fluctuate. To gain some understanding of how the SNR changes through the duration of a musical mixture, Figure 10 and Figure 11 demonstrate the effects for test track 5 at 10 × 512 averaging, which performed particularly well. It is clear that both parts largely maintain a SNR around 20dB when vocal and instrumental are mixed together, and significant improvements when these parts are isolated (0-10s and 30-35s).

V. Conclusions and Future Work

The findings from this work demonstrate that is it possible to optimise the deep neural network as used in the study ‘Deep Karaoke’ [10], by reducing the size of the mask used to train it. The extent to which this reduction can be applied depends on the end user requirements, and whether instrumental or vocal separation is required. In the majority of cases however, a 50% reduction in time and 50% reduction in frequencies provides an acceptable result. The original neural network was of size $2 \times (20500)^2$. A reduction as suggested would produce a network of size $2 \times (5120)^2$. Given that the computational time for training the original network was approximately one month, this reduced network would be able to complete training in just a few days on the same processor.

The next stage is to look at how these optimised masks behave in a real neural network. Several issues could arise, including the network not having enough data to be able to accurately recognise patterns, and as such taking an infinite amount of time to converge. Future work should also look at improving the averaging method used on the ideal masks. Currently this is a very crude approach, leaving orphan frequencies in some cases. This can be significantly improved by better mathematical functions and overlapping, and it is hoped that this alone would add significant improvement to the results. It is also noted that the method of comparison (SNR) used in determining the results is not ideal. This can clearly be seen where better SNR results are derived from the binary mask analysis which subjectively sounds worse. Many studies in the area of source separation use multiple, more specific types of SNR to categorise their generated samples, including signal to distortion ratio (SDR), signal to interference ratio (SIR) and signal to artefact ratio (SAR). It would be interesting to further analyse the generated samples using these measures, in order to more accurately derive a measure of performance.

Acknowledgment

We would like to thank the authors of [10] for clarification on some of the details of their work and [14] for the audio files without which this work would not have been possible.

References

[1] Cherry, E. Colin. ”Some experiments on the recognition of speech, with one and with two ears.” The Journal of the acoustical society of America 25, no. 5 (1953): 975-979.

[2] McDermott, Josh H. ”The cocktail party problem.” Current Biology 19, no. 22 (2009): R1024-R1027.

[3] Jin, Yu Gwang, and Nam Soo Kim. ”On detecting target acoustic signals based on non-negative matrix factorization.” IEICE Transactions on information and Systems 93, no. 4 (2010): 922-925.

[4] Grais, Emad M., Mehmet Umut Sen, and Hakan Erdogan. ”Deep neural networks for single channel source separation.” In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 3734-3738. IEEE, 2014

[5] Kang, Tae Gyoon, Kisoo Kwon, Jong Won Shin, and Nam Soo Kim. ”NMF-based target source separation using deep neural network.” IEEE Signal Processing Letters 22, no. 2 (2015): 229-233.

[6] Hu, Guoning, and DeLiang Wang. ”Speech segregation based on pitch tracking and amplitude modulation.” In Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp. 79-82. IEEE, 2001.

[7] Li, Yipeng, and DeLiang Wang. ”On the optimality of ideal binary timefrequency masks.” Speech Communication 51, no. 3 (2009): 230-239.

[8] Madhu, Nilesh, Ann Spriet, Sofie Jansen, Raphael Koning, and Jan Wouters. ”The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses.” IEEE Transactions on Audio, Speech, and Language Processing 21, no. 1 (2013): 63-72.

[9] Hartmann, William, and Eric Fosler-Lussier. ”Investigations into the incorporation of the ideal binary mask in ASR.” In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 4804-4807. IEEE, 2011.

[10] Simpson, Andrew JR, Gerard Roma, and Mark D. Plumbley. ”Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network.” In International Conference on Latent Variable Analysis and Signal Separation, pp. 429-436. Springer International Publishing, 2015.

[11] Simpson, A., 2016. Private communication

[12] Grais, Emad M., and Hakan Erdogan. ”Single channel speech music separation using nonnegative matrix factorization and spectral masks.” In Digital Signal Processing (DSP), 2011 17th International Conference on, pp. 1-6. IEEE, 2011.

[13] Emiya, Valentin, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. ”Subjective and objective quality assessment of audio source separation.” IEEE Transactions on Audio, Speech, and Language Processing 19, no. 7 (2011): 2046-2057.

[14] Bittner, Rachel M., Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. ”MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research.” In ISMIR, vol. 14, pp. 155-160. 2014