PERCEPTUAL QUALITY OF AUDIO SEPARATED USING SIGMOIDAL MASKS

Toby Stokes\textsuperscript{1}, Chris Hummersone\textsuperscript{1}, Tim Brookes\textsuperscript{1} and Andrew Mason\textsuperscript{2}

\textsuperscript{1}Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK
\textsuperscript{2}BBC Research and Development, Centre House, 56 Wood Lane, London, W12 7SB, UK

Correspondence should be addressed to Toby Stokes (t.stokes@surrey.ac.uk)

ABSTRACT
Separation of underdetermined audio mixtures is often performed in the Time-Frequency (TF) domain by masking each TF element according to its target-to-mixture ratio. This work uses sigmoidal functions to map the target-to-mixture ratio to mask values. The series of functions used encompasses the ratio mask and an approximation of the binary mask. Mixtures are chosen to represent a range of different amounts of TF overlap, then separated and evaluated using objective measures. PEASS results show improved interferer suppression and artifact scores can be achieved using softer masking than that applied by binary or ratio masks. The improvement in these scores gives an improved overall perceptual score; this observation is repeated at multiple TF resolutions.

1. INTRODUCTION AND BACKGROUND
Separation of sounds that overlap in time and frequency is often performed using a Time-Frequency (TF) mask [1]. The TF mask applies a weighting to each element of a spectro-temporal representation of the audio mixture. In general, the regions corresponding to energy from the target signal are weighted more heavily than those which correspond to energy from the interfering signal. This weighted TF representation is then resynthesised to produce the separated audio.

Separating audio in the TF domain is advantageous as it can introduce sparsity, which can aid solving underdetermined source separation problems [2]. Having acquired estimates of the TF target and interferer signal—usually by either independent component analysis [3], non-negative matrix factorisation [4] or computational auditory scene analysis [5]—these estimates can be used to resynthesise audio by masking the TF representation according to the target-to-mixture ratio within each TF cell.

Much discussion exists surrounding how the TF target and interferer estimates should be used to inform the separation and resynthesis of the target audio. The switching function, which
maps the target-to-mixture ratio to a mask coefficient, determines how much target-to-mixture based discrimination is applied when separating the mixture.

Wang [6] makes the case for using a binary switch for separation. The binary mask is formulated at each TF cell as

$$M_{IBM} = \begin{cases} 
1 & \text{if } X > Y \\
0 & \text{Otherwise}
\end{cases}$$

(1)

where $X$ represents the magnitude of the energy due to the target source and $Y$ the magnitude of the energy due to the interfering source. This provides the maximum possible discrimination between TF elements that are mostly target and those that are mostly interferer.

A commonly suggested alternative to the binary mask is the ratio mask [7] which is formulated,

$$M_{IRM} = \frac{X}{X + Y}$$

(2)

meaning each coefficient is the ratio of the target energy to total energy in the corresponding TF element. A comparison of the two found the ratio mask to be superior in terms of SNR but the binary mask to be easier to calculate [8].

A study using sigmoidal masks finds them to provide better separation, in terms of source-to-distortion ratio (SDR), than either the binary or ratio masks [9]. The formulation of the sigmoidal mask is similar to that of the ratio mask,

$$M_{SM} = \frac{X^p}{X^p + Y^p}$$

(3)

with each term raised to a power, $p$. The value of $p$ changes the S-shape of the mask’s switching function so that:

- as $p \to \infty$ the mask becomes binary;
- at $p = 1$ the mask is a ratio mask; and
- $p = 0$ produces a mask that is 0.5 at all values.

These properties can be used to create a range of masks which include the ratio mask and an approximation of the binary mask. The range used in this experiment will be centered around $p = 1$ and use exponentially scaled $p$ values to cover a wide range of possible switching functions.

Time-Frequency masking has been noted to introduce artifacts into the separated audio. This was first observed with binary masks [10] but also occurs to a lesser extent with ratio masking [11]. The artifacts might be a result simply of the effect of applying rapidly changing envelopes (in the time or frequency domain) to signals which, when created naturally, do not generally have such sharp transitions. Indeed, there is evidence that signals that have been decomposed in the frequency domain need particular care when processing, in order to avoid undesirable artifacts such as the ‘birdies’ associated with low-bitrate MPEG Audio Layer II coding [12]. Work to ameliorate artifacts includes smoothing in the cepstral domain [13], fine-shifting and adding the mask [10], as well as adding noise to the mask [14]. These previous studies point to a trade-off between artifacts and interferer suppression.

The aim of this paper is to determine the artifact-interferer trade-off across the range of sigmoidal switching functions and to quantify the effect on the change in overall audio quality. This is achieved by separating multiple audio mixtures using TF masks, created using a range of sigmoidal switching functions, and comparing the separated audio.

This work relates to a number of previous studies including work on sigmoidal switching functions [9], which is expanded here by the inclusion of masks with $p$ values between zero and one, and by the use of a more comprehensive range of metrics. The design of the range of sigmoidal masks allows it to incorporate the ideal ratio mask (IRM) and an approximation of the ideal binary mask (IBM); this allows the work to contribute to the discussion surrounding the benefits of these two masks [8, 11].

The rest of this paper is structured as follows: Section 2 describes the experimental method detailing the creation of the audio mixtures, calculations of the separated audio and the metrics used to compare the results. Section 3 gives analysis of the results. Section 4 extends the analysis of the overall perceptual score to multiple TF resolutions. Section 5 concludes the paper and summarises the main findings.
2. METHOD

This experiment was completed in three steps: firstly, audio mixtures were created and their TF overlap calculated to create a corpus containing an evenly distributed range of overlaps; secondly, for each mixture, an ideal mask was calculated using the known pre-mixture audio and each of the sigmoidal switching functions; and, finally, the audio was separated and then evaluated using the perceptual evaluation for audio source separation (PEASS) metrics [15, 16]. The software implementation was the same as that used for [14]. This section describes the processes involved in each of the three steps.

2.1 Audio Mixtures and Overlap Calculation

The 22 audio mixtures were generated from 10 second audio files with a sample rate of 24 kHz. Target signals were speech from a radio broadcast and SQAM [17]. Interferer signals were a range of background sound effects and ecological noise from the CHiME corpus [18]. Twenty-two combinations of target and interferer were selected from a larger pool of mixtures according to how much they were deemed to overlap in the TF domain.

The TF overlap was measured using analysis of the histogram of the IRM. The ratio mask gives a good indication of overlap; in each element the extreme ratios, zero and one, indicate that there is no overlap between the sources whereas the central ratio, 0.5, indicates that the sources are entirely overlapping. This idea is the basis of the overlap metric used in this work.

To calculate the overlap, an 8192-point short-time Fourier transform was performed on the target and interferer signals. The IRM of these signals was then calculated and the following process was performed for the TF locations where the target signal exceeded the -96 dBFS noise floor of the 16-bit signals.

Firstly, the IRM was calculated as in (2), then an eleven-bin histogram was calculated from the elements of $M_{IRM}$,

$$ h = \text{hist}_{11}(M_{IRM}) $$

Next, $h$ was weighted in proportion to the amount of overlap represented by each bin. The sixth (middle) bin of the histogram contains IRM elements with values near 0.5, the maximum overlap; this was weighted at one. Either side of this mid point the weighting decreased linearly and symmetrically until reaching zero at bins one and eleven,

$$ w = [0, 0.2 \ldots 1, 0.8 \ldots 0] $$

Finally, the weighted histogram was summed and divided by $n$, the number of target elements exceeding the 16-bit noise floor, to produce the final measurement,

$$ o = \frac{1}{n} \sum_{i=1}^{11} h_i w_i $$

The above process measures the spread of the histogram of the ratio mask. While this could also have been achieved using kurtosis the method employed had two distinct advantages. Firstly, kurtosis is calculated about the mean of the data; this means that two histograms with different means but similar shapes would have had similar overlap scores. The weighting of the metric used in this study was centred about 0.5 ensuring only the most severe overlap received the highest rating. Secondly, it has been noted that “kurtosis for bimodal distributions is not necessarily negative” [19]. The histograms which represent little overlap are bimodal and a kurtosis based metric would have been difficult to interpret.

2.2 Estimates

Estimates of the target audio were generated in three steps: firstly, for each of the target, interferer and mixture a cochleagram was created. Secondly, the target and interferer cochleagrams were used to create the sigmoidal mask according to (3). Finally, the mask was applied to the mixture cochleagram to obtain the estimate of the target audio.

The cochleagrams were generated using the process in [5]. Each cochleagram was made using a bank of 128 fourth-order gammatone filters spaced on the equivalent rectangular bandwidth scale up to 12 kHz, the Nyquist frequency. The cochleagram was then generated from the gammatone filterbank using a rectangular window of length 320 ms and a 50% overlap.

To generate the range of sigmoidal masks. The $p$ value in (3) was varied to produce the series of different functions. Initially, 11 masks were used with the value of $p$ scaled exponentially such that it takes values from the series of powers of two in

$$ p = q^{rac{i-1}{N-1}} $$

where $q$ is the base of the exponent, the number of masks $N$, and $i$ is the index of the mask.
the range $2^{-5}$ and $2^5$. These sigmoids are shown in Figure 1.

Each of the chosen mixtures was separated using ideal sigmoidal masks at each $p$ value. To allow comparison of the switching functions, analysis of the results was performed across mixtures at each $p$ value.

### 2.3. Metrics

Results were measured using the PEASS toolkit [15, 16]. The toolkit allows modelling of four subjective metrics: the target perceptual score (TPS), the interferer perceptual score (IPS), the artifact perceptual score (APS) and the overall perceptual score (OPS). These four metrics build a picture of the subjective quality of a separation with the final metric providing an overall score. These metrics are consistent with those used in [14], [11] and [20]. The PEASS toolkit also includes BSS Eval objective metrics [21], which are useful in understanding whether an improved PEASS score is the result of an improvement in an underlying physical metric or purely perceptual.

PEASS metrics have been used previously to compare source separation systems from a number of researchers; systems using the binary mask had a low APS and correspondingly a poor OPS [20]. The PEASS metric has also been used previously to compare the binary and ratio masks [11]. The ratio mask was found to provide superior performance but neither mask achieved an OPS greater than 40 on the 100-point scale.

### 3. PEASS ANALYSIS

The separated audio obtained in the previous section was analysed using the PEASS toolbox. The results were collated across the range of $p$ values used so the mean effect of changing the sigmoid could be analysed. Figure 2 shows the mean and 95% confidence intervals for each PEASS metric.

The results generally showed expected behaviour: changing the switching function changed the amount of discrimination between TF elements, based on the proportion of target energy they contained, and led to a worsening of artifacts, when a large amount of discrimination was applied, and to low interferer suppression, for low amounts of discrimination. This
trade-off led to little variation in the OPS across large parts of the sigmoid range.

The area where the plot does not obey the artefact-interference trade-off is around the $p$ value $2^{-1}$. At this point the IPS was higher than at any of the values when $p \geq 2^0$. This, combined with the APS being above the low values it took when $p \geq 2^0$, gave a strong peak in the OPS scores. Due to the interesting results in this region, further results were generated at $p$ values in intervals of $2^{\frac{1}{3}}$ between $2^{-3}$ and $2^{0}$. The maximum improvement in the OPS, recorded at $p = 2^{-\frac{4}{3}}$, was a full 38 points over the IRM ($p = 2^0$) and 49 points over the IBM (approximated at $p = 2^5$).

The BSS Eval metrics, shown in Figure 3, gave similar results with the SIR plateauing near $p = 2^{-1}$. The highest SAR value, while the SIR was unchanging, occurred at $2^{-1}$ giving the optimised point for the artefact-interferer trade-off.

Figure 4 shows the effect of TF Overlap on the location and height of the peak value for each mixtures at all $p$ values. The correlation between overlap and the peak value appears to be negative: as the overlap increased the peak OPS decreased. There appears to be little effect on the location of the peak: as the amount of overlap changed the peak value remained centred around $p = 2^{-1}$. This may change at higher overlaps but further data would be required to investigate this.

4. RESOLUTION

This section seeks to determine whether the results obtained previously are an effect of the TF resolution used. The artefact-interferer trade-off identified in both this work and previous studies may be affected by the TF resolution used for the analysis. The amount of switching that takes place is directly related to the window length as shorter windows entail more transitions. Conversely, the suppression of the interferer is reliant on short windows to allow localised attenuation of interfering sources.

To determine if the resolution has an effect, the OPS was measured at sixteen different TF resolutions using four of the sigmoidal values from the previous Figures 3 and 4.
Fig. 5: The change in OPS with varying sigmoids, window length and number of filters.

section: the flat, optimal, ratio and approximately-binary masks, found at the sigmoidal $p$ values of two to the powers of $-5$, $-\frac{5}{3}$, 0 and 5 respectively. The time resolution was altered by changing the window length and the frequency resolution was changed by varying the number of gammatone filters. The number of filters was changed between 32, 64, 128 and 256, and the window length took the values 80, 160, 320 and 640 samples. The results of this work are shown in Figure 5. This shows that there is little variation in OPS due to the change in time or frequency resolution and that the optimum found in the previous stage of this study is still optimal at all resolutions tested. The sigmoid defined by $p = -\frac{5}{3}$ gave the peak OPS value of 70 at all frequency resolutions and similar values were obtained across the range of window lengths.

5. SUMMARY AND CONCLUSION

This work has determined the artifact-interferer trade-off across sigmoidal switching functions, with $p$ values ranging from $2^{-5}$ to $2^5$, and quantified its effect on the overall audio quality. This was achieved using a range of 17 sigmoidal switching functions to generate ideal TF masks for audio separation. Twenty-two audio mixtures, selected to provide a range of different TF overlaps, were separated using these masks.

Where the switching function produces a near-binary mask ($p = 2^5$) the interferer suppression is good but the artifact performance is poor. Where the switching function is almost flat ($p = 2^{-5}$) the artifact performance is very good but there is little interferer suppression. The peak OPS results, recorded at $p = 2^{-\frac{5}{3}}$, exceed those provided by the IRM, by 38 points, and an approximation of the IBM, by 49 points. The changes in the perceptual scores are supported by changes in the physical metrics and are repeated at 16 different TF resolutions. Discussion in previous literature has focussed on the optimality of either the binary mask or the ratio mask. This study suggests that neither of these masks is perceptually optimal.

The amount of TF overlap appears to be negatively correlated with the maximum OPS value obtainable. However, at least for mixtures having an overlap of 0.7 or less, the degree of overlap has little effect on the optimal sigmoid.

ACKNOWLEDGEMENTS

This research is funded by the EPSRC and the BBC.

6. REFERENCES


