# THE THRESHOLD OF PERCEPTUAL SIGNIFICANCE FOR TV SOUNDTRACKS

*Robert J. Acheson and Trevor R. Agus*

SARC
Queen's University, Belfast
Belfast, Northern Ireland
racheson04@qub.ac.uk | t.agus@qub.ac.uk

## ABSTRACT

Hearing loss affects 1.5 billion people world-wide [1], affecting many aspects of life, including the ability to hear the television. Simply increasing the volume may restore audibility of the quietest elements, but at a cost of making other elements undesirably loud. Therefore, at the very least, dynamic range compression could also be useful, fitted to an individual's frequency-dependent hearing loss. However, it is not clear whether the audibility of the quietest parts of TV audio needs to be preserved. This experiment aims to measure which elements of the audio are important by presenting normal-hearing listeners with binary masked versions of TV audio presented at 60 dB(A), muting audio below a given sensation level. It was hypothesised that spectro-temporal regions with the most power density would dominate perception, such that the less active regions may not be missed. To find this threshold of perceptual significance, a two-alternative forced choice signal detection experiment was designed in which excerpts from BBC television shows were binary masked and presented to the participants, with the task to identify which clips sounded more processed. The results suggest that discarding audio below 10 phons would rarely be noticed by most listeners.

## 1. INTRODUCTION

This paper is part of a larger project which aims to improve the listening experience for people with hearing loss when watching the television. Hearing loss affects one in five people in the UK [2]. The most widely adopted device for improving hearing is the behind-the-ear (BTE) hearing aid [3] however, it has inherent limitations since it must work in real time with a low power computer. Another approach to improving audibility is the use of clean audio for multimedia [4] i.e., audio which has all sounds except dialogue attenuated. However, most clean audio solutions consist of algorithms used to remove or attenuate background noise—including music and sound effects, which are fundamental elements of television soundtracks which contribute to the narrative of television programmes.

### 1.1. Hearing Aids

There are two main categories of hearing loss: conductive and sensorineural [5]. Both hearing losses cause a raised threshold of audibility. In order to restore audibility of quiet sounds, they must be amplified, although loud sounds cannot be amplified as this would cause discomfort for the person with hearing loss. Hearing aids provide multi-band dynamic-range compression in order to restore audibility in the affected frequency ranges to improve audibility, without amplifying sounds beyond the level of discomfort [6]. Due to their small size, only a small battery can be fitted to hearing aids. This means the potential computational power of the hearing aid is limited, in order to preserve battery life. The hearing aid is also required to work in real time; due to its low computational power and impossibility of a "look ahead" feature, i.e., the ability to view the oncoming amplitude envelope, the juxtaposition of quiet and loud sounds can often impair audibility due to over-correction, often referred to as overshoot [7]. For example, if someone were to clap their hands, the hearing aid would attenuate the frequency ranges in which the clap is present in order to prevent damage to the user's ears. However, the slow release-time of the hearing aid means that any conversation following the clap may also be attenuated for a brief time.

Hearing aids are limited in ways which affect both the frequency domain and the temporal domain. The frequency bands on which the multi-channel dynamic-range compression is applied are broad, resulting in low frequency resolution. Temporal resolution limitations are caused by the necessity of real-time operation in the form of latency. With a view to improve on the hearing aid for pre-recorded entertainment purposes, the audio can be processed offline. This means that the frequency and temporal resolutions can be greatly improved upon, and the main limitation is computation time. However, considering that the audio is pre-recorded, computation time is less of a pressing issue.

## 2. PROPOSED SOLUTION

The proposed offline process is based on Ray Meddis' BioAid algorithm, which is used in [8]. BioAid is similar to the proposed solution in that it uses a filter bank at the beginning and end of the signal chain and is psychoacoustically inspired, however the filter banks used in the proposed solution are comprised of auditory gammatone filters rather than Butterworth filters. The proposed solution to be used in this project takes the form of a program designed in MATLAB and consists of two main parts: the analysis tool and the adaption tool, see figure 1. The analysis tool is used to examine the incoming audio and compare the amplitudes of frequency bands to a predefined threshold. It can then generate an adapted amplitude envelope for each frequency band based on the input amplitude. The adaption tool uses the amplitude envelopes generated by the analysis tool to adjust the amplitude envelopes of the frequency bands. Presumably, the audio broadcast on television or streaming services has been mixed such that the important elements are audible to people with normal hearing. Furthermore, if a person with normal hearing cannot tell when some of the au-

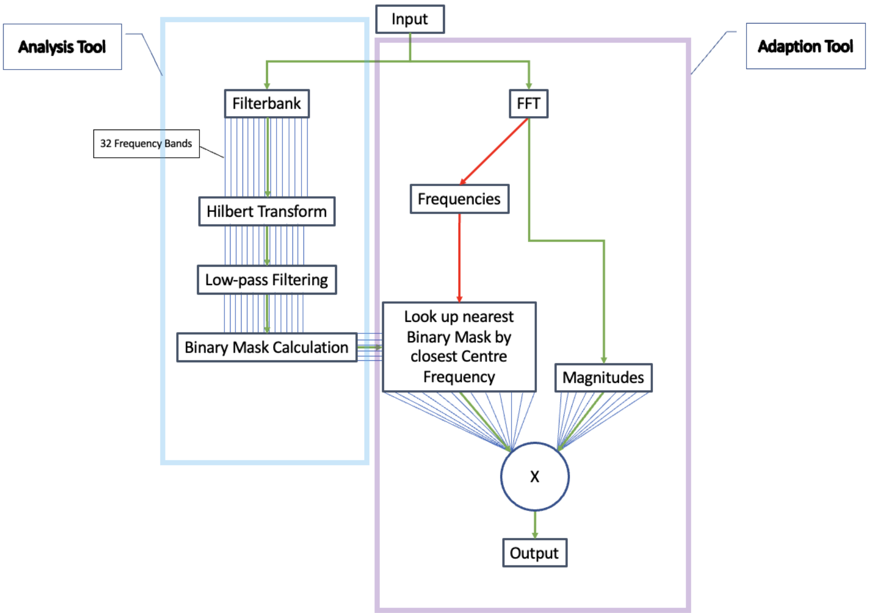

Figure 1: *Signal Flow Chart for the Analysis and Adaption Tool, see sections 2.1-2.3.*

dio has been removed, then it could be deemed unimportant to the listening experience. Using this logic, the important elements of the audio can be identified by muting parts of the dynamic range and testing to see if normal hearing people can hear a difference.

## 2.1. Analysis Tool

To analyse the audio from "the ear's perspective", we use a gammatone auditory filterbank. The analysis tool takes an audio file as the input and filters it into frequency bands using the gammatone filterbank. This auditory filter bank is comprised of $n$ band-pass gammatone filters, which are modelled on the filters found in the human cochlea [9]. This means that the audio is separated in the program with similar frequency resolution to the cochlea. After filtering the audio into frequency bands, the amplitude envelope of each frequency band is calculated in order to compare it to the given threshold.

### *2.1.1. Filter Bank*

The gammatone filters used in the filterbank are modelled on the human cochlea, designed by Hohmann [9] and implemented using the Auditory Modelling Tool- box 1.2 [10]. The bandwidths of these filters are based on experimental data (Equivalent Rectangular Bandwidth, ERBs). These bandwidths are closely related to the critical bandwidth of the auditory filters found in the human cochlea and were measured using the notched noise method [11]. The notched noise method involves measuring the audibility threshold of a signal in a masker, where the signal is a fixed frequency sine tone, and the masker is noise with a notch centred on the signal frequency. The ERB of a filter can approximated with the following formula, see equation 1 [12]. To avoid phase delays and group delays, the audio is filtered twice, the second filtering being time reversed such that any phase or group delays caused by the filters are corrected. To maintain the bandwidths of the original filterbank, the bandwidths of the original filters are doubled.

$$ERB(f) = 0.108f + 24.7 \tag{1}$$

The filterbank takes a mono input and has parallel outputs, one for each of $n$ filters. Each of the frequency bands can be denoted by the centre frequency ($f_c$) of the gammatone filter used to create them.

### *2.1.2. Amplitude Envelope Measurement*

The amplitude envelope of each frequency band is calculated to provide a means of comparing the input signal to the normal hearing (NH) audibility thresholds. The amplitude envelope of each channel is measured here using the Hilbert transform. This is an efficient way to measure the amplitude envelope for signals resembling sinusoids as used in [13], [14], and [15].

### *2.1.3. Low Pass Filtering*

The fastest amplitude modulations from the output of the Hilbert transform were removed by a low-pass filter. As seen in Figure 2, the magnitude of the Hilbert transform contains the overall amplitude envelope of the input signal and a rectified version of the temporal fine structure. The low pass filter reduces all frequencies over half the bandwidth of the frequency band leaving a smoother amplitude envelope for the fitting algorithm to compare the input signal to the given threshold.

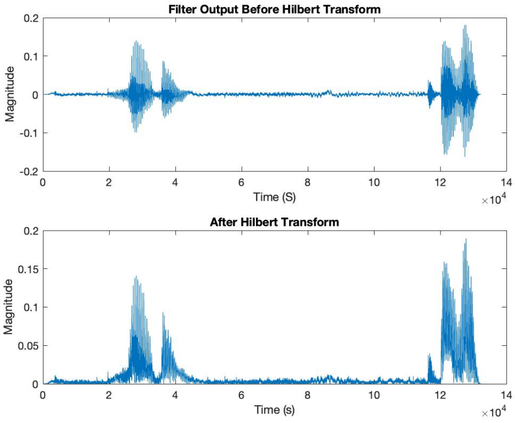

Figure 2: *Amplitude Rectification Using Hilbert Transform*

### 2.2. Binary Mask

To test what of the audio above the threshold of audibility is perceived by a person with normal hearing, the smooth binary algorithm was created. Inspired by binary masking used in image processing, as used in [16], the objective is to superimpose a mask of ones and zeros onto the audio such that each amplitude below a given threshold is assigned the value zero and each amplitude above the threshold is assigned the value one. Each amplitude with an assigned value of zero can then be muted. By moving this threshold and testing whether a perceptual difference has been detected, the "threshold of perceptual significance" (TPS) can be measured. This TPS is the point above the threshold of audibility at which changes to the audio are perceptually significant. The audio below the measured TPS can be permanently muted resulting in faster computation times without changing the perception of the audio in its original form.

Some considerations taken when designing the binary masks were the naturally fast amplitude modulations, measurement of the "threshold of perceptual significance", and ensuring no latency is present.

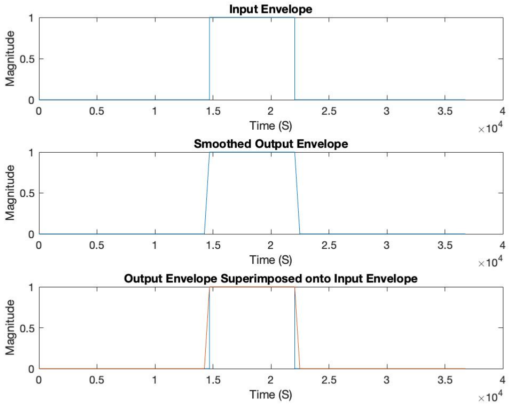

Figure 3: *Plots Showing Amplitude Ramping to Smooth Binary Mask Function*

The binary mask has fast amplitude modulation by nature; however, this leads to the creation of artifacts which are not desirable when processing audio. To remedy this, linear amplitude ramping was employed to smooth the binary mask envelope and prevent extremely fast changes in amplitude (see Figure 3). The ramping is configured in such a way that should an attack and release ramp overlap, the maximum of the two ramps is chosen (see Figure 4).

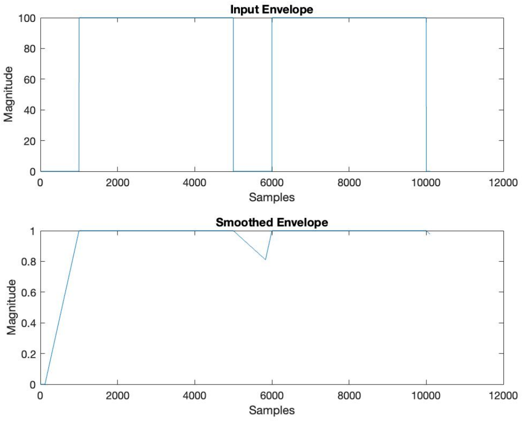

Figure 4: *Binary Mask Ramps with Overlapping Release and Attack*

Due to the fact that humans don't perceive loudness on a linear scale, the binary mask is measured in phons. This means that the threshold for the binary mask algorithm will be situated at the same perceptual loudness level across all frequencies using the ISO 226 Equal-Loudness-Level Contour model by Jeff Tackett [17].

### 2.3. Adaption Tool

The adaption tool performs a Fast Fourier Transform (FFT) on the input audio file to break it up into its comprising sine tones which can be described in terms of frequency, magnitude, and phase. It then performs a "Slow Inverse Fourier Transform" (SIFT). This SIFT re-synthesises the audio by synthesising each of the comprising sine tones individually, allowing for fine amplitude control at each frequency over time. This fine control is achieved because each sine wave can be discretely amplitude modulated up to speeds matching the sample rate of the audio. The binary mask calculated previously is used to alter the amplitude of the input audio frequency components. It should be noted that to maintain intelligibility of speech, the amplitude modulation should not be excessively fast such that the speech amplitude envelopes are distorted. Similarly, large changes should not be made to consecutive frequencies to prevent artefacts such as ringing. The formula for the SIFT is as seen in Equation 2:

$$y = \frac{1}{N}\sum_{n=0}^{N} x[f_c, n]e^{-2\pi j f n}.g[f_c, n] \tag{2}$$

where $g$ is the binary mask, $N$ is the total number of samples, $x$ is the input sample, and $y$ is the output sample.

## 3. METHODS

### 3.1. Participants

Participants for the experiment were recruited through email. Five participants with normal hearing were recruited (three male and two female), aged from 20–24, with a mean age of 22. To confirm their normal hearing, an audiometer was used to test 250 Hz–8 kHz down to 10 phons.

### 3.2. Comparison Procedure

For the duration of the experiment, the participant was seated in a sound-treated booth with a window. A monitor was placed outside the window, visible to the participant. The participants were given verbal instruction before the trials. The experiment took the form of a series of two-alternate forced choice trials. In each trial, the participant was presented two clips of audio through a pair of Sennheiser HD600 headphones, each lasting three seconds: one with binary mask processing, which will be referred to as processed; and one without binary mask processing, which will be referred to as unprocessed. The duration of the onset and offset smoothing ramps for the binary were set to 10ms and 50ms respectively. Both signals were low-pass filtered as the ISO 226 model used does not support thresholds above 12.5kHz. The processed and unprocessed clips presented in each trial were not necessarily from the same source. The order of the processed and unprocessed clips for each trial was randomised. The processed audio clips were selected randomly from a pool of audio clips for a given threshold measured in phons. The thresholds were spaced equally in intervals of five phons (5–35 phons). The participant was asked to select which of the two audio clips sounded more processed using '[1]'and '[2]'on a computer keyboard to select the first and second audio clips respectively. Their response to each trial was recorded.

### 3.3. Stimuli

The audio was sourced from BBC iPlayer. Seven television show genres were chosen from the iPlayer menu (Drama, News, Music, Documentary, Sport, Comedy, and Entertainment) and two television programs were picked at random from each genre using a random number generator (with the exception of Drama, for which only one was chosen). Three audio clips were recorded from each television show with randomised start times, each lasting five minutes. The audio used in the experiment was generated using these five-minute-long excerpts. For each of the excerpts a clip of three seconds in length was extracted with a randomised start time. This was repeated for each threshold level in phons to provide a bank of 39 audio clips per threshold level.

## 4. RESULTS

Figure 5 shows the averaged participant responses fitted to a cumulative normal distribution model. As hypothesised, the TPS is above the auditory thresholds, at approximately 17.1 phons based on a 75% threshold. The lowest level of performance at 51.7% indicates little noticeable difference between the processed and unprocessed clips. The peak performance at 92% indicates an obvious difference between the processed and unprocessed clips.

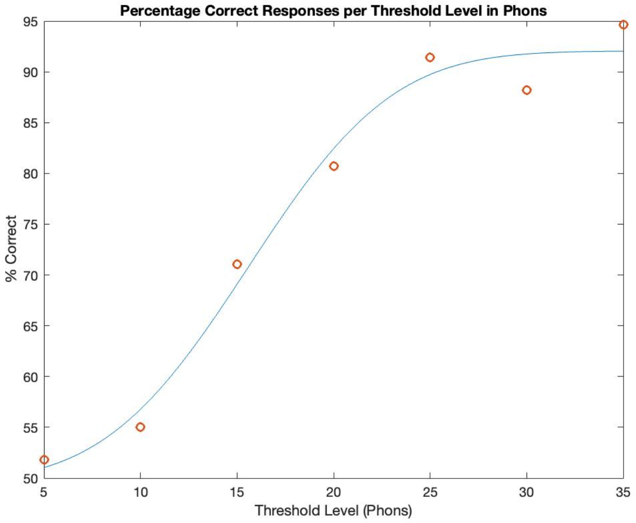

Figure 5: *The average accuracy of responses across all participants (red circles) and a cumulative normal curve fitted by the least-squares method (blue line).*

## 5. DISCUSSION

The results are consistent with the hypothesis that the spectro-temporal regions with less power can often be removed with little detriment to the listening experience. This suggests that the whole of the dynamic range does not need to be preserved at all times. Should this be the case, dynamic range compression algorithms used to adapt audio for people with hearing loss are not required to be as aggressive, which in turn reduces the risk of introducing audible artifacts.

In the current implementation of this experiment, a presentation level of 60 dB(A) was used, which returned a TPS of 17.1 phons however, it is not clear whether varying the presentation level would change the TPS. This could be investigated directly in a future experiment. Measuring the TPS for a range of presentation levels would potentially allow for automation of the binary mask threshold based on the presentation level, i.e. the TV volume in a real-world scenario.

When watching TV in real-world conditions, there would be more background noise and other distractions present compared to the experiment conditions. This could mean that the TPS may, in fact, be higher than shown in the results for real-world scenarios. With the absence of this background noise, it is possible that the participants are able to identify the processed clips due to the addition of silence as opposed to the absence of important elements of audio.

## 6. CONCLUSION

The focus of this experiment was to identify the key elements to people with normal hearing of television audio by muting parts of the audio hypothesised to be unimportant to the listening experience. The results show that at a presentation level of 60 dB(A), most audio under 10 phons could potentially be muted without affecting the listening experience. This means that not all of the dynamic range needs to be preserved at all times, allowing for less aggressive dynamic range compression algorithms, and therefore a reduced risk of introducing audible artifacts. Potential areas of future research have been highlighted in section 5 to improve the binary mask algorithm's flexibility in terms of presentation level.

## 7. REFERENCES

[1] World Health Organization, "Deafness and hearing loss," Available at https://www.who.int/health-topics/hearing-loss.

[2] RNID, "Facts and figures," Available at https://rnid.org.uk/about-us/research-and-policy/facts-and-figures/.

[3] NHS, "Hearing aids and implants," Available at https://www.nhs.uk/live-well/healthy-body/hearing-aids/.

[4] B. Shirley and P. Kendrick, "The clean audio project: Digital TV as assistive technology," 2006.

[5] Starkey, "Types and causes of hearing loss," Available at https://www.starkey.com/hearing-loss/types-and-causes.

[6] K. Patel and Issa M. S. Panahi, "Frequency-based multi-band adaptive compression for hearing aid application," *The Journal of the Acoustic Society*, vol. 146, no. 4, pp. 2959–2959, 2019, Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7380331/.

[7] Brian C. J. Moore, "Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids," *Ear and Hearing*, vol. 17, no. 2, pp. 133–161, 1996.

[8] R. Meddis, W. Lecluyse, N. R. Clarke, and T. Jurgens, "Exploration of a physiologically-inspired hearing- aid algorithm using a computer model mimicking impaired hearing," *International Journal of Audiology*, vol. 55, pp. 346–357, 2016.

[9] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acust.*, vol. 88, pp. 433–442, 2002.

[10] P. Majdak, C. Hollomey, and Robert Baumgartner, "Amt 1.2: A toolbox for reproducible research in auditory modeling," *Acta Acust.*, vol. 6, pp. 19, 2022.

[11] B. C. J. Moore and B. R. Glasberg, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1–2, pp. 103–188, 1990.

[12] Julius O. Smith and Jonathan S. Abel, "Equivalent rectangular bandwidth," Available at https://ccrma.stanford.edu/ jos/bbt/Equivalent$_{R}ectangular_{B}andwidth.html$.

[13] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. C. J. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Available at https://www.pnas.org/doi/pdf/10.1073/pnas.0607364103.

[14] G. Gilbert, M. A. Akeroyd, and S. Gatehouse, "Discrimination of release time constants in hearing-aid compressors," *International Journal of Audiology*, vol. 47, no. 4, pp. 189–198, 2008.

[15] J. C. Ziegler, C. Pech-Georgel, F. George, and C. Lorenzi, "Speech-perception-in-noise deficits in dyslexia," *Developmental Science*, vol. 12, no. 5, pp. 732–745, 2009.

[16] Jinyang Liang, Sih-Ying Wu, Rudolph N Kohn Jr, Michael F Becker, and Daniel J Heinzen, "Grayscale laser image formation using a programmable binary mask," *Optical Engineering*, vol. 51, no. 10, pp. 108201–108201, 2012.

[17] Jeff Tackett, "Iso 226 equal-loudness-level contour signal," Available at https://uk.mathworks.com/matlabcentral/fileexchange/7028-iso-226-equal-loudness-level-contour-signal.