IMPROVED AUDITORY MASKING MODELS

M. R. Flax *, E. Ambikairajah +, W. H. Holmes +, J. S. Jin *

* School of Computer Science and Engineering
+ School of Electrical and Telecommunications Engineering
University of New South Wales Australia

Abstract

Hybrid masking models are defined and are informally judged to be improved masking models. Moore's method for deriving spreading functions, a previously unused constituent, is closer to real world observations than its counterparts. Moore's method lowers the computational complexity of the masking model and suggests that people are auditorily more sensitive to high frequencies then previously assumed. The hybrid models are independent of cochlea mapping functions, hence the same model may be used for assessing auditory redundancy in a variety of mammals. The two best masking models are chosen and classed as models which a] preserve auditory quality for the loss of spectral character discrimination and b] discriminate spectral character for the loss of auditory quality.

1 INTRODUCTION

Definition of perceptually relevant information in audio signals is currently frequency threshold based. Masking thresholds are found from a two stage model. Stage one defines a frequency spreading function, stage two defines spreading function combination methods. The combination of these two stages produces a threshold below which information is deemed perceptually redundant. Masking models are most commonly used in compression systems [Sen et al. (1997), Sen et al. (1994), Black et al. (1995), Schroeder et al. (1979), Virag, N. (1995), Veldhuis et al. (1989), Johnston, J.D. (1989)] but also serve purpose in objective quality assessment systems [Beerends et al. (1992)], auditory pitch evaluation [Terhardt, E. (1979)] as well as other complex audio analysis systems. Compression systems simply ignore redundant signal information to some degree and hence are capable of compressing a signal into fewer bits for transmission. Auditory pitch evaluation requires a masking model for initial assessment of the harmonic pitch associated with a section of audio. Objective quality assessment systems use masking models, as well as other perceptually relevant transforms to define a signal's character and hence try to compare signals in a perceptually relevant domain for people.

All stages in a masking model rely on a perceptual filter bank (PFB) which imitates the nature of the mapping from frequency to cochlea location sensitivity. Such cochlea mappings outline how to place filter central frequencies and are defined for a large range of mammals [Greenwood, D.D. (1996)]. PFBs are defined for people in FIR [Glasberg et al. (1990), Moore et al. (1987), Moore et al. (1997), Moore et al. (1983)] and IIR [Slaney, M. (1993), Irino, T. (1995), Irino et al. (1998), Irino et al. (1999)] forms. All stages of a masking model work on sum-mated outputs from a PFB, called an excitation function.

To complete the introduction, current methods in both stages of masking models are reviewed, sections 1.1 and 1.2. Section 2 outlines hybrid masking models by defining each stage in each model. Section 3 reviews the results of the hybrid models and general conclusions are addressed in section 4.

1.1 The frequency spreading function - Stage One

Spreading functions describe how the spectral energy of a signal is absorbed by certain sections of the cochlea. [Lynch et al. (1997)] reviewed in detail the spreading function methods of [Terhardt, E. (1979), Schroeder et al. (1979), Veldhuis et al. (1989), Black et al. (1995)].

As an example, [Beerends et al. (1992)] developed a spreading function method based on those of [Terhardt, E. (1979)]. The generic method for [Terhardt, E. (1979)] based spreading functions is to split on the bark scale all PFB centre frequencies into scale locations which are lower and higher then

	Lower	Higher
Terhardt	$S_1 = 27$	$S_2 = 24 + 0.23(f_\upsilon/kHz)^{-1} - 0.2L_\upsilon/dB$
Beerends	$S_1 = 31$	$S_2 = 22 + min(\frac{230}{f_\upsilon}, 10) - 0.2L_\upsilon$

Table 1: Terhardt style spreading function slope (Units in dB / Bark), where f_υ is the centre frequency and L_υ is the level in dB of the reference filter (masker).

Method	Excitation threshold functions	Method	Masking threshold functions	Author
1a	$\Delta S_\upsilon = Y_\upsilon - S_\upsilon$	1b	$T_i = \sum_{\upsilon=1}^{i-1} \Delta S_\upsilon + \sum_{\upsilon=i+1}^{N} \Delta S_\upsilon$	Terhardt
2a	$\Delta S_\upsilon = Y_\upsilon - S_\upsilon$	2b	$T_i = \left(\sum_{\upsilon=1}^{N} \Delta S_\upsilon^\alpha\right)^{\frac{1}{\alpha}}$	Beerends et al.
3a	$\Delta S_\upsilon = V_\upsilon + S_\upsilon$	3b	$T_i = \sum_{\upsilon=1}^{N} \Delta S_\upsilon$	Black et al.
4a	$\Delta S_\upsilon = T_{max}(f_m) S_\upsilon$	4b	$T_i = \sum_{\upsilon=1}^{N} \Delta S_\upsilon$	Veldhuis et al.
5a	$\Delta S_\upsilon = \sum_{\upsilon=1}^{N} Y S_{-\upsilon}$	5b	$T_i = \omega_i \Delta S_i$	Schroeder et al.

Table 2: Excitation and masking threshold functions. Where N specifies the number of filters in the PFB, Y is the excitation function of the PFB, Y_υ is the excitation for the υ'th filter in the PFB, S_υ specifies the υ'th component of an N component spreading function, ΔS_υ specifies the υ'th excitation threshold of an N component excitation threshold function, $S_{-i,\upsilon}$ is the spreading function which is frequency reversed $(-i)$ and cyclicly shifted by υ frequency samples and $S_{-\upsilon}$ is the same as $S_{-i,\upsilon}$ however assumes that frequency reversal is implicit, i.e. $-\upsilon \equiv -i, \upsilon$. T_{max} is the masking threshold at the masker's frequency f_m and V_υ is a self masking threshold. T_i is the masking threshold intensity for the i'th filter of the PFB, $\alpha < 1$ and ω is a sensitivity function.

the reference filter (masker) central frequency. For either side of the masker, slopes are defined and used to specify the spreading function, refer to Table 1. These spreading functions are depicted in Figure 2. This figure suggests that Terhardt's spreading function is of larger magnitude then Beerends's at all frequencies, it is the higher curve. It slopes less quickly for all frequencies.

Of all spreading functions, [Schroeder et al. (1979)] and [Veldhuis et al. (1989)] do not define masker level dependent spreading functions.

1.2 Method of Combining the Spreading Functions - Stage 2

Masking models define a frequency based threshold below which components are deemed to be masked or redundant in the perception of an audio signal. Masking models work with spreading functions as they asses what magnitude a frequency must have to be perceived above its spread energy. Masking thresholds are derived from excitation thresholds. Excitation thresholds are derived using spreading functions and operations on either the excitation function, a masker self spread function or by multiplicative procedures, treated in Table 2. Masking thresholds are derived from operations on the excitation threshold function to arrive at a masking threshold function, treated in Table 2. The results of these steps, the masking threshold, is the output of a masking model and is depicted in Figure 1.

2 PROPOSED HYBRID MASKING MODEL

A hybrid masking model is built by defining two stages. Stage one, the frequency spreading function is outlined in section 2.1 and replaces those of section 1.1. Suitable spreading function combination methods are chosen in section 2.2 from section 1.2 and used as stage two.

2.1 The frequency spreading function

[Moore et al. (1983)] defines a novel method for deriving frequency spreading functions. They call such spreading functions excitation patterns and actually derive spreading functions for each filter in the PFB independently. Essentially spread energy from one filter is found by assessing how much energy leaks to all of the other filters in the PFB. Hence the magnitude at the $\upsilon'th$ filter's central frequency

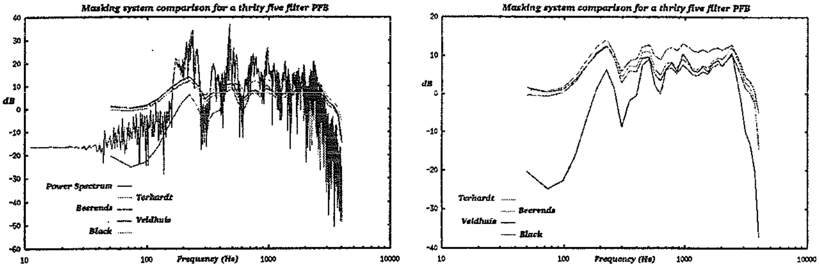

Figure 1: Comparison of masking models for a frame of female speech. Left: Depicted against the power spectrum. Right: Depicted alone are (from the 300 Hz vertical - bottom to top) the masking models by [Veldhuis et al. (1989), Terhardt, E. (1979), Beerends et al. (1992), Black et al. (1995)].

	Hybrid Masking Models	
	Spreading function	Excitation / Masking Threshold Methods
i	Moore	Black (Methods 3a and 3b)
ii	Moore	Terhardt (Methods 1a and 1b)
iii	Moore	Beerends (Methods 2a and 2b)

Table 3: Hybrid masking model definition

(f_v) is found for each filter (i) of the PFB and aligned with the central frequency of the filter they were generated by (f_i). This is defined

$$S_v(f_i) = O_i(f_v) \tag{1}$$

where $S_v(f)$ is the magnitude of spreading function for the $v'th$ filter of the PFB at frequency f, O is the output of the PFB and $O_i(f)$ is the magnitude of the $i'th$ filter of the PFB at the frequency f. The derivation of the spreading functions have a constant computational complexity, $Order(c)$ for some constant c.

Figure 2 shows examples of [Moore et al. (1983)] based spreading functions. As the spreading functions are dependent on the audio information alone, they are different for every audio frame in an audio stream. Slopes generated by equation 1 are less mechanical in appearance then the spreading functions previously defined.

2.2 Method of Combining the Spreading Functions

Spreading functions are derived for each filter in the PFB. For this reason the chosen combination method must operate on unique spreading functions, the combination methods proposed by [Schroeder et al. (197 Veldhuis et al. (1989)] are not applicable.

2.3 Hybrid Masking Models

Suitable spreading function combination methods of those reviewed in section 1.2, will be combined with the frequency spreading function specified in section 2.1 to yield the masking models defined in Table 3.

It is important to note that hybrid models ii and iii are independent of cochlea mapping functions, such as bark, critical band or ERB scales. This is not the case for hybrid model i, which depends on the bark scale for assessing the self masking level. The implications of this are that any PFB may use hybrid models ii and iii and not have problems if the pitch to cochlea mapping functions are altered. Such mapping functions are implicit in the placement of PFB filter centre frequencies. These same models may be applied to an arbitrary PFB which has arbitrary filter transfer functions, which implies application to arbitrary mammalian auditory systems.

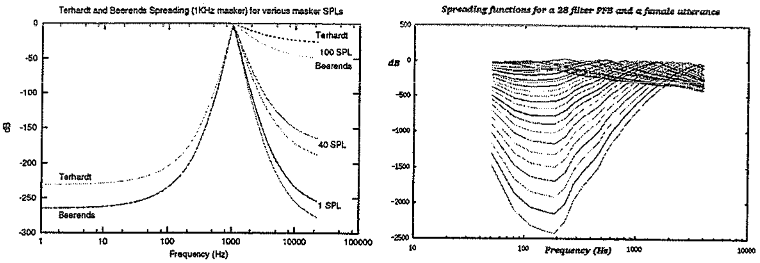

Figure 2: Left: Terhardt / Beerends spreading function example. Right: Moore spreading functions derived by a twenty eight filter PFB for a female spoke phrase.

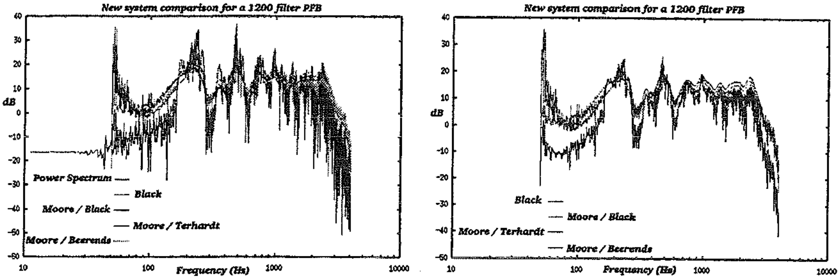

Figure 3: The three hybrid masking models specified in section 2.3 after processing the first frame of a spoken phrase. Depicted against the power spectrum (Top) and alone (Bottom).

3 RESULTS AND DISCUSSION

A female phrase, split into three frames, is filtered through a PFB. Each of the three hybrid masking models specified in section 2.3 are applied to the spreading functions.

Figure 3 depicts the results of the new masking models, as well as the model of [Black et al. (1995)]. For high frequencies (above 3 kHz), hybrid models ii and iii suggest a lower masking threshold then the Black and hybrid i models. For mid frequencies (200 Hz to 3 kHz) all of the models propose a similar masking threshold. For low frequencies, all models except hybrid i agree to within around three decibels on the masking threshold, hybrid i assumes a lower threshold.

For all frequencies, hybrid model ii follows the signal power spectrum closely. It has peaks and troughs which greatly exceed the other masking models in magnitude. Discrimination of power spectrum peaks and troughs, the character of the audio signal, is preserved by all hybrid models except hybrid ii.

3.1 Redundancy analysis

Consider the redundancy assumed by each model, we may quote redundancy as the percentage of Fourier frequency bins which are below the masking threshold. Table 4 lists such redundancies frame by frame and the average over all frames. Interestingly this table suggests that hybrid model i assumes less redundancy then the original Black model. The model with the least redundancy is hybrid ii and the most is the Black model. Ranking the models which yield the most average redundancy to least we have ; Black, Hybrid iii, Hybrid i, Hybrid ii.

Model	Frame 1	Frame 2	Frame 3	Average
Black	86%	85%	87%	86%
Hybrid i	81%	78%	71%	77%
Hybrid ii	35%	37%	44%	39%
Hybrid iii	74%	80%	91%	82%

Table 4: Fourier bin redundancy comparison.

	Hearing Researchers				General Public			
Rank	Black	Hybrid i	Hybrid ii	Hybrid iii	Black	Hybrid i	Hybrid ii	Hybrid iii
1'st	0/22	0/22	22/22	0/22	0/29	1/29	28/29	0/29
2'nd	0/22	6/22	0/22	16/22	0/29	12/29	1/29	16/29
3'rd	1/22	16/22	0/22	5/22	2/29	16/29	0/29	11/29
4'th	21/22	0/22	0/22	1/22	27/29	0/29	0/29	2/29

Table 5: Informal listening test results.

3.2 Listening test results and discussion

An informal listening test was conducted where hearing researchers and people of the general public were invited to listen to a speech phrase and four reconstructions of the same phrase with perceptually redundant Fourier frequency bins removed. The test is deemed informal as the subjects were all using different equipment as the test was distributed using the Internet. The four reconstructions corresponded to the four masking models of Black, hybrid models i, ii and iii. 15% (9 of the total 60) participants gave inconclusive opinions, as they failed to rank all of the reconstructed phrases. As a large majority of participants gave conclusive results, the test was considered valid.

Hearing researchers subjectively suggest that the perceptual quality ranking is (Table 5) : Hybrid model ii, Hybrid model iii, Hybrid model i, Black. The general public suggest the following perceptual quality ranking (Table 5) : Hybrid model ii, Hybrid model iii, Hybrid model i, Black. From the tabulated results we may draw the following implications. The best masking model is the hybrid model ii. The worst model is the Black model. Hybrid model iii performs better then hybrid model i.

The results imply that the second best (according to the listening test) masking model is also the model with the second largest redundancy level. This is an interesting fact, as it suggests that the theory of redundancy in audio signals is a reality. This conclusion is drawn as subjective quality ranking does not match objective redundancy ranking.

Hybrid masking model ii yields perceptually the best quality whilst removing redundancy, however its functional shape follows the audio signal's power spectrum extremely closely. This suggests that on the macro level this model does not highlight power spectrum artifacts such as formants of speech or simply power spectrum peaks and troughs. It is therefor suggested that masking models be classified as models which discriminate quality for the loss of character and models which discriminate character for the loss of quality.

4 CONCLUSIONS

To date masking models have used synthetic frequency spreading functions which are purely stationary with a signals intensity or varying with a signals intensity to derive masking thresholds. Assessing redundancy in an audio signal may be improved by using real world frequency spreading functions such as those suggested by [Moore et al. (1983)]. An implication of [Moore et al. (1983)]'s method for deriving spreading functions is that computational complexity becomes constant, as opposed to previously linear computational complexities.

The resulting masking model has the potential to operate as a black box. Some of the hybrid masking models given within this paper are such models. These hybrid models confirm the mid frequency sensitivity prediction of peoples auditory systems by [Black et al. (1995)], however suggest greater sensitivity at high frequencies then previously predicted.

When considering an audio stream's redundancy versus perceptual quality, different models are suited to different tasks. Hybrid masking model ii is convincingly the best of new masking models for auditory processing which requires high quality for the loss of auditory features, such as speech coders. Hybrid masking model iii slightly outperforms hybrid masking model i for auditory processing which requires auditory features for the loss of perceptual quality, such as auditory pitch assessment.

References

Beerends J.G., Stemerdink J.A. (1992) "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation", J. Audio Eng. Soc., 40, 963-978

Black M., Zeytinoglu M. (1995) "Computationally efficient wavelet packet coding of wide-band stereo audio signals", ICASSP, 5, 3075 -3078

Glasberg B.R., Moore B.C.J. (1990) "Derivation of auditory filter shapes from notched-noise data", Hearing Research, 47, 103-138

Greenwood, D.D. (1996) "Comparing octaves, frequency ranges, and cochlea-map curvature across species", Hearing Research, 94

Irino, T. (1995) "An Optimal Auditory Filter", IEEE Workshop on App. of Sig. Proc. to Audio and Acous., 198-201

Irino T., Unoki M. (1998) "A time-varying analysis/synthesis auditory filterbank using the gammachirp", ICASSP98 , 3653-3656

Irino T., Unoki M. (1999) "An analysis/Synthesis Auditory FilterBank Based on an IIR Implementation of the Gammachirp", JASJ

Johnston, J.D. (1989) "Perceptual transform coding of wideband stereo signals", ICASSP89, 1993 -1996

Lynch M., Ambikairajah E., Davis A. (1997) "Comparison of Auditory Masking Models for Speech Coding", EuroSpeech-97, 3, 1495 - 1498

Moore B.C.J., Glasberg B.R. (1983) "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", J. Acous. Soc. Am., 74, 750-753

Moore B.C.J., Glasberg B.R. (1987) "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns", Hear. Res., 28, 209-25

Moore B.C.J., Glasberg B.R. and Baer T. (1997) "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness", J. Audio Eng. Soc., 45, 224-40.

Schroeder M.R., Atal B.S., Hall J.L. (1979) "Optimizing digital speech coders by exploiting masking properties of the human ear", J. Acous. Soc. Am., 66(6), 1647-1652

Sen, D., Holmes, W.H. (1994) "Perceptual enhancement of CELP speech coders", ICASSP-94, 105-108

Sen, D., Allen, J.B. (1997) "A new auditory masking model for speech and audio coders", Speech Coding For Telecomm. Proc., 53 -54

Slaney, M. (1993) "An Efficient implementation of the Patterson-Holdsworth auditory Filter Bank", Apple Comp. Tech. Rep. #35

Terhardt, E. (1979) "Calculating Virtual Pitch", Hear. Res., 1, 155-182

Veldhuis R.N.J., Breeuwer M., Van Der Waal R. (1989) "Subband coding of digital audio signals without loss of quality", ICASSP, 3, 2009 -2012

Virag, N. (1995) "Speech enhancement based on masking properties of the auditory system", ICASSP, 1, 796 -799