Proceedings of the Institute of Acoustics

USING AN AUDITORY MODEL TO DETERMINE WHEN SPEECH IS MASKED BY NOISE

Paul Murrin(1), David M. Howard(1), Andy M. Tyrrell(1) and Paul Barrett(2)

(1) Department of Electronics, University of York, Heslington, York, YO10 5DD.
(2) BT Laboratories, Martlesham Heath, Ipswich, Suffolk, IP5 3RE.

1. INTRODUCTION

Due to the way our hearing system analyzes sounds, a quiet sound will become inaudible in the presence of a louder noise - a process termed masking. This paper describes a method for predicting which parts of one sound are masked by another when both are incident at a human ear. As an example, consider the masking effect of a vehicle passing a mobile telephone user. Some parts of the speech signal become completely inaudible due to the presence of the background noise.

An auditory model converts the masking sound to a time-frequency representation with an equivalent resolution to that of our hearing system. From this representation a masked threshold is produced below which another sound will be inaudible. A similar auditory model produces an excitation pattern for the sound being masked. Figure 1 shows an auditory spectrogram (excitation) for a speech signal, figure 2 shows the long term average excitation and masking threshold of a noise signal used to mask the speech. The excitation pattern and masked threshold are compared over all time-frequency elements and those areas of the excitation pattern which are below the masking threshold are labeled as masked. This process will result in a masking decision spectrogram as shown by figure 3. If all frequencies at a given time are masked then the entire signal is denoted as being masked for that time interval. Early analysis has been performed to determine the proportion of speech that is masked by common environmental noises at different signal-to-noise ratios. Results from an informal listening test show that no degradation is observed when masked speech is removed.

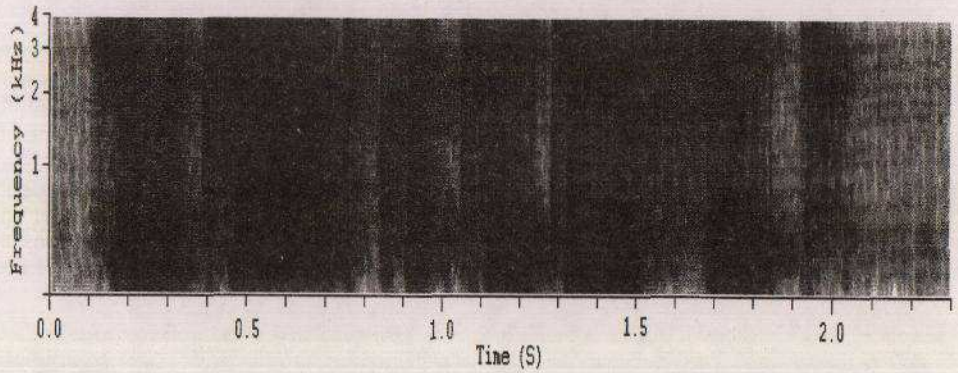

Figure 1: SPL scaled speech excitation ("He retired quickly to his seat." adult female, active speech level -26dBov).

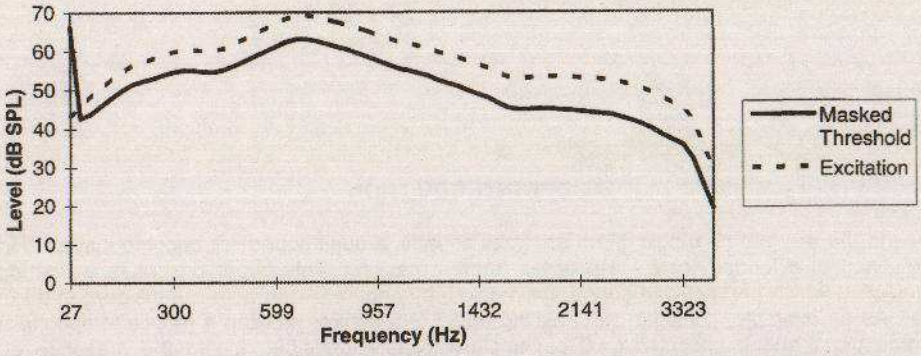

Figure 2: Average excitation and masking threshold of street noise at -26dBov RMS).

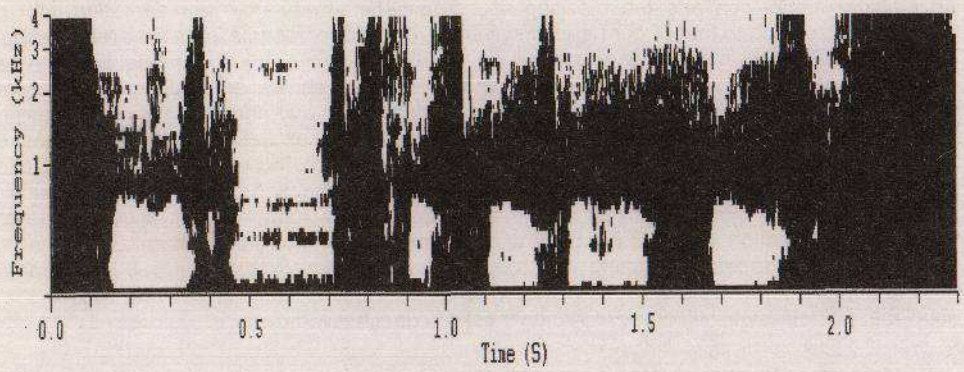

Figure 3: Binary masked decision spectrogram (white = audible, black = masked).

2.THE AUDITORY MODEL

The auditory model consists primarily of a bank of 64 bandpass filters†. The filter shapes are chosen to be similar to those derived from psychoacoustic experiments [1]. Filters towards the lower end of the frequency range have a narrower bandwidth than those towards the higher end. The Bark domain is a frequency transformation which represents characteristic place along the human cochlea. It has been shown that auditory filters each have a similar shape in the Bark domain.

† Currently the model operates over a 4kHz bandwidth for analysis of telephone quality speech. This is easily modified to cover the full audio bandwidth.

The signal to be analyzed is filtered simultaneously by each of the auditory filters. The energy at the output of each auditory filter is obtained by averaging the squared filter outputs over a rectangular window whose width varies dependent upon the filter's bandwidth. The auditory model can produce a variety of spectrographic output formats:

1. SPL scaled spectrogram - showing a physical representation of the loudness using a similar time-frequency resolution to the ear.

2. Sone scaled spectrogram - showing a perceptual representation of the loudness using a similar time-frequency resolution to the ear.

3. Masking threshold - showing the minimum level that an additional time-frequency component must attain before it may be heard when played alongside the sound.

2.1 Generating the filters

Sixty four auditory filters are placed with their center frequencies evenly spaced along a Bark scaled axis covering a zero to 4kHz frequency range. The filters each have a similar shape in the Bark domain, this shape formed by two straight lines is shown by figure 4. Auditory filters are non-linear since the lower skirt of each filter becomes less steep with increasing input level resulting in an increase in bandwidth. This effect is modeled using time variant FIR filters.

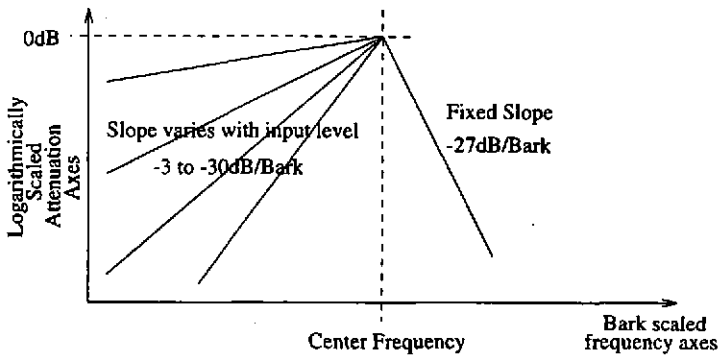

Figure 4: The filter shape in the Bark domain.

To derive a set of FIR coefficients for a given filter, the Bark domain response must first be transformed to a linear frequency domain with linear scaled attenuation axis. Equation (1) from [2] is used to convert frequency from Hertz to Bark. The frequency sampling method of digital filter design [3] is applied to obtain coefficients for a 255 tap FIR filter. These coefficients are windowed using a Kaiser window to smooth ripples in the resulting filter response. Figure 5 shows the frequency response of a number of these filters on a linear frequency domain.

$$F_{Bark} = 13\arctan(0.76F_{kHz}) + 3.5\arctan\left[\left(\frac{F_{kHz}}{7.5}\right)^2\right] \quad (1)$$

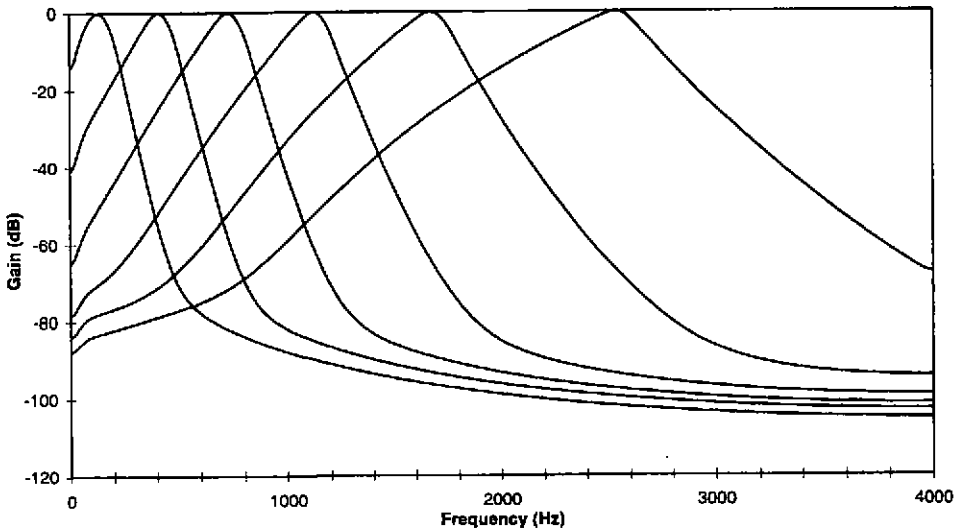

Figure 5: Measured response of a number of filters in the linear frequency domain.

2.3 Output

To obtain the envelope of the output signal's energy the signal is squared and averaged over a rectangular window. The width of the window varies depending on the bandwidth of the filter to reflect the varying temporal response of the ear (high temporal resolution at high frequencies and high frequency resolution at lower frequencies). The equivalent rectangular bandwidth (ERB) of the filter is determined and used to set the width of the averaging window as 1/(2*ERB) which provides a reasonable temporal response whilst minimizing ripple and smearing.

The model can produce a variety of output types from the filter energy. These were listed earlier and are now described in more detail.

2.3.1 Excitation

The excitation type of output yields a spectrogram which shows the time-frequency-amplitude plot of the input waveform. The time-frequency resolution is similar to that of the human hearing system due to the shape and placing of the auditory filters. When compared to typical wide and narrow band spectrograms the auditory spectrogram provides key attributes of each [6]. For speech signals, closely spaced vertical lines, or striations, due to each vocal fold closure are visible at higher frequencies (usually only visible on a wide band spectrogram) and the harmonics of voiced speech are clearly visible at the lower frequencies (usually only visible on a narrow band spectrogram).

The SPL scaled excitation spectrogram shows the physical loudness of the sound whereas the Sone scaled spectrogram yields the perceived loudness of the sound. Conversion from SPL to Sone is performed using equation (2). Note that the Sone conversion makes use of the absolute threshold of hearing. This is a frequency dependent level below which our ears cannot detect sound. The threshold is lowest around 1-5kHz, where the majority of important speech information lies, and rises sharply at higher and lower ends of the audible frequency range. Perceived loudness rises quickly near threshold and logarithmically thereafter; equation (2) simulates this effect.

$$L_{Sone} = \frac{20}{35}\left(L_{SPL} - \frac{Th(f)}{2}\right) - 50e^{-0.1(L_{SPL}-Th(f))} + C \qquad (2)$$

The threshold *Th(f)* is found using Therhardt's approximation (equation (3) from [4]) which is also used in the MPEG-1 audio coding scheme.

$$Th(f)_{dB} = 3.64 f_{kHz}^{-0.8} - 6.5e^{-0.6(f_{kHz}-3.3)^2} + 10^{-3} f_{kHz}^{4} \qquad (3)$$

The sone scaled spectrogram yields a much clearer view of the harmonics and formants of voiced speech than the SPL scaled spectrogram by applying the threshold of hearing and logarithmic growth of loudness. Much of the 'image noise' in the SPL scaled auditory spectrogram is suppressed whilst important features tend to be emphasized for speech signals.

2.3.2 Masked Threshold

When an audio signal is presented to the ear that signal has the effect of changing the threshold below which other audio signals cannot be heard. The masked threshold output from the auditory model aims to predict this modified threshold. The masked threshold is always lower than the excitation level produced by the signal. The difference between the excitation level and masked threshold is dependent both upon the frequency and on the noise-like or tone-like nature of the masking signal. Equations (4),(5) and (6) are used to determine the masked threshold level [5].

$$Mth_{dB} = L_{dBSPL} - Oi_{dB} \qquad (4)$$

$$Oi_{dB} = \alpha(14.5 + f_{Bark}) + (1-\alpha)a_v \qquad (5)$$

$$a_v = -2.0 - 2.05\arctan\left(\frac{f_{kHz}}{4}\right) - 0.75\arctan\left(\frac{f_{kHz}^{2}}{2.56}\right) \qquad (6)$$

The noise-like or tone-like nature of the input signal (coefficient of tonality, α in equation (5)) is determined using a spectral flatness measure derived from a measure of the spectral entropy.

3. DETERMINING MASKED SPEECH

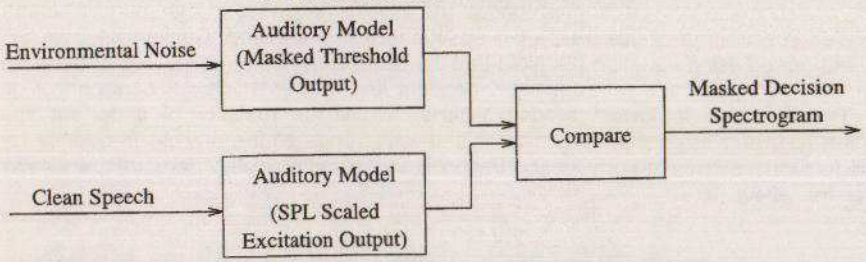

Figure 6: System for determining parts of one audio signal masked by another.

Figure 6 shows how the auditory model is used to determine which components of one audio signal are masked by another. The masking signal is applied to the auditory model and a masked threshold is output. The masked signal is also applied to the auditory model and an SPL scaled excitation spectrogram is output. Each time-frequency component of the two auditory model outputs are compared and a decision as to whether each component is masked can be made. The results are shown as a binary masking decision spectrogram. Figure 7 shows the masking effect of a car passing a speaking male.

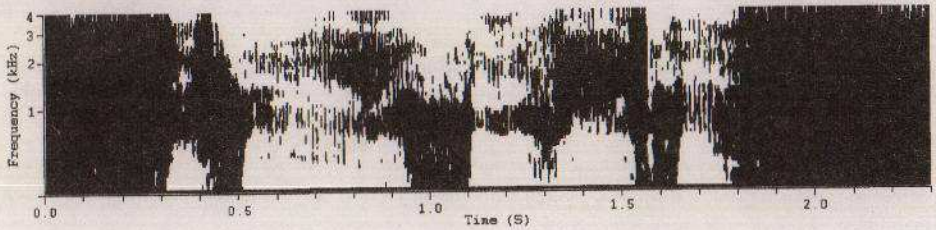

Figure 7: Binary masked decision spectrogram (white = audible, black = masked).

Alternatively the signals level above the masked threshold may be plotted as shown by figure 8.

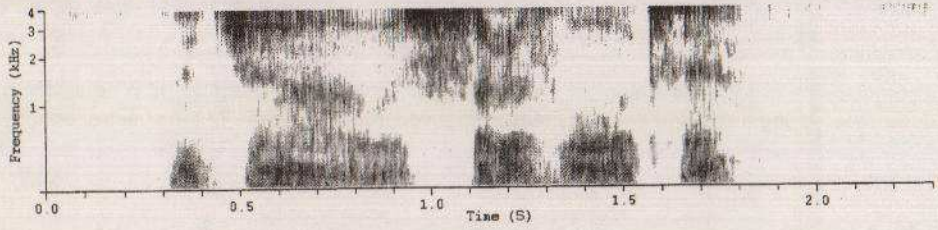

Figure 8: Signal level above masked threshold (white = inaudible).

One proposed application of the identification of masked speech is to assist with the objective assessment of voice activity detectors (VADs). In this situation the data required is the temporal locations of those parts of speech which are completely masked at all frequencies over a certain time period. This can be easily obtained from the binary masking decision spectrogram. The following table shows the percentage of active speech which is masked by various different noise types at a number of signal-to-noise ratios.

Noise type	SNR (dB)	Male Speech	Female Speech	Child Speech
Babble	+10	1.46%	1.11%	1.40%
	+5	4.31%	4.38%	4.30%
	0	9.13%	10.21%	7.87%
	-5	18.04%	18.13%	15.84%
Vehicle	+10	1.02%	0.07%	0.41%
	+5	2.48%	1.25%	1.50%
	0	5.62%	6.04%	3.47%
	-5	10.45%	14.03%	7.25%
Street	+10	1.10%	1.11%	1.76%
	+5	3.43%	3.33%	3.11%
	0	8.18%	7.29%	6.63%
	-5	15.85%	15.35%	14.08%

Table 1: Percentage of speech masked for various speakers and noise types.

As expected, it can be seen from the table that as the signal-to-noise ratio increases the percentage of the speech signal which is masked decreases. The data in the table indicates that only a small percentage of the speech signal is completely masked. It is likely that more of the signal is masked but this method is very harsh at making the masked decision. For example if the excitation at one frequency band is just a fraction of a dB above the masking threshold then the signal is marked as unmasked for that time interval even though it is unlikely that the signal will be detected. As one example, only 4.31% of the male speech is marked as masked by babble noise at an SNR of +5dB even though 53% of the time-frequency elements of the speech signal are below the masked threshold.

To evaluate the effectiveness of the method of detecting masked speech the masking decision can be used to gate the clean speech before mixing it with the noise. The resultant signal should sound identical to the original since the speech which has been removed would be inaudible due to masking by the noise. A listening test was devised to test this hypothesis. The subjects were asked to compare two versions of each utterance against a reference and decide which was closest to the reference. Ideally they should not be able to hear any difference. The results show this to be the case.

Within the test a number of cases were used to examine the effects of relaxing the masking decision by evenly raising the masking threshold with a given offset. An interesting result was that with a 10dB rise in threshold the clipping of the speech was audible by careful listening but the test results indicated that the clipping was not detected. This effect is probably a result of the listening test design. A further listening test of a "Degradation Category Rating" (DCR) type will be performed to study this further.

4. CONCLUSIONS AND FURTHER WORK

A method for determining the masking effect of one audio signal over another has been described, and results presented which show the masking effect occurring. An early assessment of the ability of the proposed system to detect masked speech has been performed by gating the clean speech using the masking decision and mixing this with environmental noise. The listening test shows that the auditory model is capable of detecting parts of an audio signal which are masked by background noise.

One problem with this approach for detecting the masking of speech by background noise is that mixing clean speech and environmental noise together does not produce a true representation of the speech produced in noisy conditions. A number of changes in speech parameters have been identified when speaking in noisy conditions. Probably the most obvious of these is the Lombard effect discovered in 1911 which is the "spontaneous tendency of speakers to increase their vocal intensity when talking in the presence of noise"[7]. Other changes include [8][9]: raising the pitch of speech, using a greater proportion of voiced speech to unvoiced speech, reduction of speaking rate and adjusting the spectral tilt of our speech.

To overcome this problem the speech could be recorded by a subject wearing headphones with the required background noise played over the headphones. Care must be taken to ensure that the correct level of sidetone feedback is available so that the subject may talk naturally as if they were in the noisy environment. Work is underway to study the effect of speech spoken in noise in the context of the masking effect of the noise.

5. ACKNOWLEDGMENTS

This work is jointly funded by the EPSRC and British Telecommunications PLC.

6. REFERENCES

[1] Moore B.C.J., "*Hearing: Handbook of perception and cognition*", 2nd Ed., London: Academic Press, 1995

[2] Colomes C., Lever M., Rault J.B., Dehery Y.F., Faucon G., "*A Perceptual Model Applied to Audio Bit-Rate Reduction*", J. Audio Eng. Soc., Vol. 43, No. 4, pp. 233-239, April 1995

[3] Ifeachor E.C. & Jervis B.W., "*Digital Signal Processing: A Practical Approach*", Addison-Wesley Publishers Ltd., 1993

[4] Therhardt E, "*Calculating Virtual Pitch*", Hearing Research, Vol. 1, pp. 155-182, 1979

[5] Kapust R, "*A Human Ear Related Objective Measurement Technique Yields Audible Error and Error Margin*", AES 11th Int'l Conf., pp191-202, May 1992

[6] Brookes T.S., "*A Real-Time Auditory Spectrograph*", D.Phil. Thesis, York University, 1996

[7] Pick H.L. Jr., Siegel G.M., Fox P.W., Garber S.R., Kearney J.K., "*Inhibiting the Lombard effect*", J. Acoust. Soc. Am., Vol. 85, No. 2, pp. 894-900, February 1989

[8] Hanley T.D. & Steer M.D., "*Effect of Level of Distracting Noise upon Speaking Rate, Duration and Intensity*", Journal of Speech and Hearing Disorders, Vol. 14, pp363-368, (1949)

[9] Pisoni D.B., Bernacki R.H., Nusbaum H.C., Yuchtman M., "*Some Acoustic-Phonetic Correlates of Speech Produced in Noise*", Proc. IEEE ICASSP '85, pp1581-1584, 1985