FITTING THE MEL SCALE

S. Umesh
Dept of Electrical Eng., I.I.T, Kanpur-208016, India.

L. Cohen
Dept. of Physics, City University of New York, New York, NY 10021, USA.

D. Nelson
U.S. Dept. of Defense, Ft. Meade, MD 20755, USA.

ABSTRACT

We show that there are many qualitatively different equations, each with few parameters, that fit the experimentally obtained Mel scale. We investigate the often made remark that there are two regions to the Mel scale, the first region ($<\sim$ 1000 Hz.) being linear and the upper region being logarithmic. We show that there is no evidence, based on the experimental data points, that there are two qualitatively different regions or that the lower region is linear and upper region logarithmic. In fact $F_M = f/(af+b)$ where F_M and f are the mel and physical frequency respectively, fits better then a line in the linear region or a logarithm in the "log" region.

1. INTRODUCTION

The Mel scale is a fundamental result of psychoacoustics, relating real frequency to perceived frequency. The foundation of the mel scale is the classic work of Stevens and Volkman [5]. Their results and variations thereof appear in almost every speech book. In this paper we address certain issues regarding the Mel scale and in particlar we discuss the issue of fitting the Mel scale and the implications and physical meaning of the fitting formula. Many authors have attempted fitting formulas for a variety of reasons. Most important among these is that such fits sometimes give an indication of the underlying physical phenomenon. Furthermore, such fits may be used to develop models for the psychoacoustic scale. We believe that the first to attempt such an analysis was Koenig [3] although the formula by Fant [2] is the most commonly referenced. Also, O'Shaughnessy [4] has given a formula which is the same functional form as that of Fant but with different parameters.

It is very often said that the Mel scale is linear up to a point (most authors use the figure 1000 Hz.) and logarithmic after that. This is an important statement that may or may not be factual but often it is said to imply certain aspects of the physical situation. The focus of this paper is to fit curves of different functional form to the original data. We demonstrate that based on the numerical evidence solely there is insufficient evidence to justify a conclusion that there are two functionally different regions.

It appears that the first attempt to devise a formula that is acoustically relevant was by Koenig [3]. Koenig divided the physical frequency region of 0 to 10,000 Hz. in two different regions and argued that the relevant acoustic scale is linear below 1000 Hz and logarithmic beyond that. As Fant makes it clear he was motivated by Koenig and felt that his formula was better because the "discontinuity at 1000 Hz." is removed.

Numerical Data

The results presented here are based on the original Stevens and Volkman paper As we did not have access to the numerical data we read the points from the graph of Stevens and Volkman. This of course produces some errors but we believe it is accurate enough for our considerations. We also point out that the original data probably had considerable variability since they were averaged over many different listeners. We give these numbers in Table 1. In this table f represents the physical frequency and F_M for the Mel frequency.

2. FITTING CURVES

We have attempted to fit a variety of functional forms. Our choice was motivated by various ideas regarding

$f =$	40	161	200	404	693	867	1000	2022	3000	3393	4109	5526	6500	7743	12000
$F_M =$	43	257	300	514	771	928	1000	1542	2000	2142	2314	2600	2771	2914	3228

Table 1: These are the physical frequency in Hertz and the corresponding Mel values, obtained from the figure in the Stevens and Volkman paper.

speech production [6, 7]. We have found the best fits for the functional forms indicated by the tables. The error criterion was the usual sum of residuals. We have also presented the fitting formulas of Fant and O'Shaughnessy and have kept their numerical values. We note that their functional form is the same as Case 4.

We considered three frequency ranges [1]:

Table 2: The range considered in the Stevens and Volkman paper; 40 Hz. to 12000 Hz.

Table 3: The so called "linear region", 40 Hz to 1000 Hz.

Table 4: The upper region, often called the "log" region, 1000 Hz to 12000 Hz.

Generally we see that many equations fit the data quite well. Of particular interest is the simple Case 1,

$$F_M = \frac{f}{af + b} \tag{1}$$

This simple equation also fits the linear and log region quite well as can be seen from Tables 3 and 4.

3. IS THERE A LINEAR AND LOG REGION?

As mentioned in the introduction it has often been stated that the region from 0 to 1000 Hz. is linear and the region above is logarithmic. It certainly is true that it looks linear but it also looks like the beginning of many non linear functions and therefore linearity may not be an important physical condition. We believe that one must be very careful in assigning meaning to the phrase "it is linear up to 1000 Hz". We point out that there are two possible meanings to the phrase: a) it can convey that there are two qualitatively different regions based on different physical effects or b) that there may be a non-linear curve representing the physical phenomena but that it just happens to be approximately linear in that regions. For example in radioactive decay the decay law $N = N_0 e^{-\lambda t}$ is exponentially decreasing and its explanation is well grounded in theory and experiment. It does not have two qualitatively different regions. For some decay times it is approximately linear ($N(t) = N_0 - \lambda t$) - however it would be misleading to imply that there are two qualitatively different regions. We believe that many papers on speech have assumed, or imply that indeed there are two qualitatively different regions.

To study this issue we used only the points up to 1000 Hz and refit the curves as above. The results are presented in Table 3. Indeed the linear fit is quite good and one can certainly say that the data is approximately linear. However note that other curves also fit the data as well. Therefore there is no justification in saying that there are two qualitatively different regions or that there are two different physical effects for different regions (that may be the case but the only point we are trying to make is that the data does not support that view.) To study this issue further one or all of the following must be done: a) perform new experiments along the same lines; b) use the data from other type of experiments and c) perhaps develop a realistic model of the physical phenomena that would indicate whether indeed there are two qualitative regions.

Consider now the region above 1000 Hz. The log fits as given by Cases 3, 4 or 10 give a pretty good fit but they certainly are not better than the equation given by cases 1 and 2. Hence, again one can not come to any firm conclusions that there is a log region based on the data.

4. SCALING OF FREQUENCIES

There has been considerable work on attempting to find and correlate the physical reasons that would produce the Mel scale. One concept that has been developed by the authors and others is scaling of the spectrum We will not go into the details here but if we had simple (uniform) scaling then the relation would be $F_M = lnf$. However the mel scale does not follow this curve. Recently the authors have shown that one can experimen-

[1] Also, we thought that it would be interesting to fit the data only over the regions that is of most importance to speech and hence we also fitted the data in the frequency range to 161 to 7743 Hz. The results turned out to be basically the same as in Table 2 and are not presented.

Case	Equation	Constants	Error
1	$F_M = f/(af + b)$	a = 0.00024, b = 0.741	14776
2	$F_M = f/(af + blogf)$	a = 0.000218, b = 0.108	20958
3	$F_M = alog(b + f/c)$	a = 2561, b = .961, c = 616.6	43174
4	$F_M = alog(1 + f/b)$	a = 2620, b = 657.6	46391
5	$F_M = a + blnf + c(lnf)^2$	a = 1608, b = -1901, c = 574	78744
6	$F_M = a + blogf + cf$	a = - 1845, b = 964, c = 0. 1204	499852
7	$F_M = a + blnf$	a = -2978, b = 629	1276556
8	$F_M = a + bf$	a = 652, b = 0.284	2603035
9	$F_M = a + b/f + c/f^2$	a = 2306, b = -541487, c = 18097219	4736402
10	$F_M = alogf$	a = 541	7453270
11	$F_M = alog(l + f/b)$	Fant: a = 1000/log 2, b = 1000	307980
12	$F_M = alog(l + f/b)$	O'Shaughnessy: a = 2595, b = 700	112940

Table 2: Different fitting formulas for the entire range of observations in Stevens'paper,(40 Hz to 12000 Hz). Cases 11 and 12 are the formulas of Fant and O'Shaughnessy. They are of the form give by Case 3. For the sake of interest and comparison we have included the good fits we obtained and some bad fits.

Case	Equation	Constants	Error
1	$F_M = f/(af + b)$	a = 0.0004, b = 0.603	1008
2	$F_M = f/(af + blogf)$	a = 0.000168, b = 0.1188	383
3	$F_M = alog(b + f/c)$	a = 3362, b = 1.03, c = 1043	351
4	$F_M = alog(1 + f/b)$	a = 2342, b = 596	777
5	$F_M = a + blnf + c(lnf)^2$	a = 3294, b = -3080, c = 773	332
6	$F_M = a + blogf + cf$	a = -458, b = 275, c = 0.646	556
7	$F_M = a + blnf$	a = - 1859, b = 407.8	11416
8	$F_M = a + bf$	a = 127.7, b = 0.9	2341
9	$F_M = a + b/f + c/f^2$	a = 1346, b -424328, c = 4.08×10^7	8818
10	$F_M = alogf$	a = 246	284654
11	$F_M = alog(l + f/b)$	Fant: a = 1000/log 2, b = 1000	4572
12	$F_M = alog(l + f/b)$	O'Shaughnessy: a = 2595, b = 700	1256

Table 3: Curve-fitting the Mel scale for frequencies below 1000 Hz. While the linear fit is good, it can be seen that many of the non-finear equations also fit the data well in this region.

Case	Equation	Constants	Error
1	$F_M = f/(af + b)$	a = 0.000244, b = 0.773	4327
2	$F_M = f/(af + blogf)$	a = 0.000222, b = 0.1056	13763
3	$F_M = alog(b + f/c)$	a = 2131, b = -0.07, c = 341	18239
4	$F_M = alog(1 + f/b)$	a = 2520, b = 593	38747
5	$F_M = a + blnf + c(lnf)^2$	a = -6699, b = 2848, c = -98	17243
6	$F_M = a + blogf + cf$	a = -5947, b = 2303, c = -0.015	15138
7	$F_M = a + blnf$	a = -5474, b = 934	18291
8	$F_M = a + bf$	a = 1322, b = 0.19	628424
9	$F_M = a + b/f + c/f^2$	a = 3634, b = -6.04×10^6, c = 3.416×10^9	14906
10	$F_M = alogf$	a = 642	1952263
11	$F_M = alog(l + f/b)$	Fant: a = 1000/log 2, b = 1000	303230
12	$F_M = alog(l + f/b)$	O'Shaughnessy: a = 2595, b = 700	111300

Table 4: Different fitting formulas for the "LOG' range (1000 Hz to 12000 Hz).

tally obtain the Mel scale (to a high approximation) from a speech production point of view and we have also considered the implication of this scale from the point of view of scaling the spectrum. Further discussion of scaling and the Mel scale will be presented elsewhere.

5. CONCLUSION

The fact that so many curves (with few parameters) fits the data so well leads one to the conclusion that caution must be used in assigning significance to the interpretation of a particular functional representation. Since the existing data may be well fit by curves of many functional forms, there is insufficient evidence to justify the selection of one of these forms as correct. The conclusion may be made that the warping function is not linear, and it is not logarithmic, and the function may be approximated numerically, but these are the extent of the conclusions which may be drawn from the existing data.

As to the issue of different regions its clear that there is no evidence that there are two qualitatively different regions. In particular there is no evidence that the lower region is linear and upper region logarithm. In fact it can be argued that case 1 fits every region quite well. It is nether linear nor logarithmic. We would like to make clear that we are not saying that indeed these two regions are not linear or logarithmic. It may be the case that they are, based on further evidence, and the lack of a good fit may be due to errors in the data. However, we are emphasizing that based solely on the Mel scale data such conclusions are unwarranted.

Finally, we would like to point out that there are other experiments that relate response of the ear to actual physical frequencies. A review of such experiments has been given by Allen [1]. We are currently investigating that data in the light of the results of this paper.

6. ACKNOWLEDGEMENT

This work was supported by HBCU/MI Program, the Indian AICTE Career Award for Young Teachers, and the CUNY FRAP award program.

7. REFERENCES

[1] J.B. Allen and D. Sen, "A bio-mechanical model of the ear to predict Auditory Masking", *preprint,* 1998.

[2] G. Fant, "Acoustic Theory of Speech Production", *Mouton & Co.*, The Hague, 1960.

[3] W Koenig, "A new frequency scale for acoustic measurements" *Bell Telephone Laboratory Record,* vol. 27, pp. 299-301, 1949.

[4] D. O'Shaughnessy, "Speech Communication - Human and Machine" *Addison-Wesley*, New York, 1987.

[5] S. Stevens and J. Volkman, "The relation of pitch to frequency" *American Journal of Psychology*, vol. 53, P. 329,1940.

[6] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale Transform In Speech Analysis" *IEEE Transactions on Speech and Audio Processing*, Scheduled to appear in Jan 1999 issue.

[7] S. Umesh, L Cohen, and D. Nelson, "Warping Functions in Speech" *Proc. Int. Soc. Opt. Eng.,* to appear 1998.