The importance of perceptive adaptation of sound features in audio content processing

Silvia Pfeiffer
University of Mannheim, Germany
{pfeiffer@pi4.informatik.uni-mannheim.de}
Lehrstuhl für Praktische Informatik IV
University of Mannheim, Germany

Technical Report 18/98

Abstract

In analyzing audio material for features useful for extracting content, we must consider the value gained by adapting our analysis algorithms to the analysis processes of the human ear. This aspect with regard to loudness features is thoroughly examined in this paper. The increase in correlation to be gained by such cognitive processing is about 10%.

1 Introduction

Psychophysical experiments with human cognition of environmental stimuli has revealed the fact that human perception does not coincide with the physically measured intensity of the stimuli. G.T. Fechner (1801-1887) stated a logarithmic relationship now known to be not generally applicable. Relations between stimuli and cognition are a lot more difficult [ZF90].

The focus of the MoCA (movie content analysis) project at the University of Mannheim is the extraction of semantic information from video material. Semantic information from the audio track of videos is also determined. In this context, the question frequently arises whether preprocessing the digitized signals according to human perception is worth the added expense.

In the work presented here, we examine this problem with respect to loudness measures of sound signals. Loudness is the most basic information contained within an audio signal, similar to color for a video signal. In videos, for example, it may be used to determine suspense, because an increase (or decrease) in one normally accompanies an increase (or decrease) in the other. Another useful application would be to determine action based on the amount of change in loudness. Therefore, it is vital to have a measure which gives the computer exact information about the level of perceived loudness by a human at a certain instant.

We have implemented a perceptive loudness measure based on findings in psycho-acoustics. To prove the importance of such cognition-based preprocessing, we determined the correlation of our perceptive loudness measure with human judgments and compared the results to simple sound level measures.

2 Definition of loudness measures

Our cognitive adaptation process begins with the **sound intensity** $[dB]$, which is based on the amount of energy emitted from a sound source I, put into relation to the threshold of audibility of a 1 kHz tone $I_0 = 10^{-12} \frac{J}{m^2 \cdot sec}$:

$$L = 10 \log_{10} \frac{I}{I_0} \tag{1}$$

The sound intensity, however, can not be easily measured and is therefore frequently replaced by the **sound pressure level** (SPL) L $[dB]$. This defines the relation between the currently effective sound pressure p_x as exercised on a microphone membrane and the reference sound pressure $p_0 = 2 \cdot 10^{-5} \frac{N}{m^2}$, which relates to the threshold of audibility I_0:

$$L = 20 \cdot \log_{10} \frac{p_x}{p_0} \; [dB] \tag{2}$$

Both equations result in the same values and are therefore used interchangeably in this work.

When comparing sinus tones of different frequency but same SPL, we find that their perceived loudness is also different. In extensive experiments, equal-loudness level contours (so-called isophones) were determined [Mol73, ZFS57]. They define the **loundess level** L_N $[phon]$ of a tone of frequency f by giving the SPL of a tone of $1kHz$ which sounds identically loud. Figure 1 shows the isophones, which have been standardized by ISO [ISO61, ISO87].

The isophones are defined by measured SPL levels of different but fixed frequencies (see [ISO61] for the table). In order to calculate a loudness level of a sinus tone of frequency f_x and SPL L_x, we had to implement an algorithm which related the position of point (f_x, L_x) to the four closest points in the table: (f_s, L_1), (f_s, L_2), (f_l, L_3) and (f_l, L_4) with f_s being the highest frequency in the isophone table lower than f_x, f_l the lowest frequency higher than f_x, and L_1, L_2, L_3 and L_4 chosen such that (f_x, L_x) lies within the parallelogram built by $(f_s, L_1), (f_s, L_2), (f_l, L_3)$ and (f_l, L_4) (see Figure 2). Using these points, we can calculate two straight lines: (f_s, L_1) - (f_l, L_3) and (f_s, L_2) - (f_l, L_4). These straight lines approximate the isophones quite well within the regarded area.

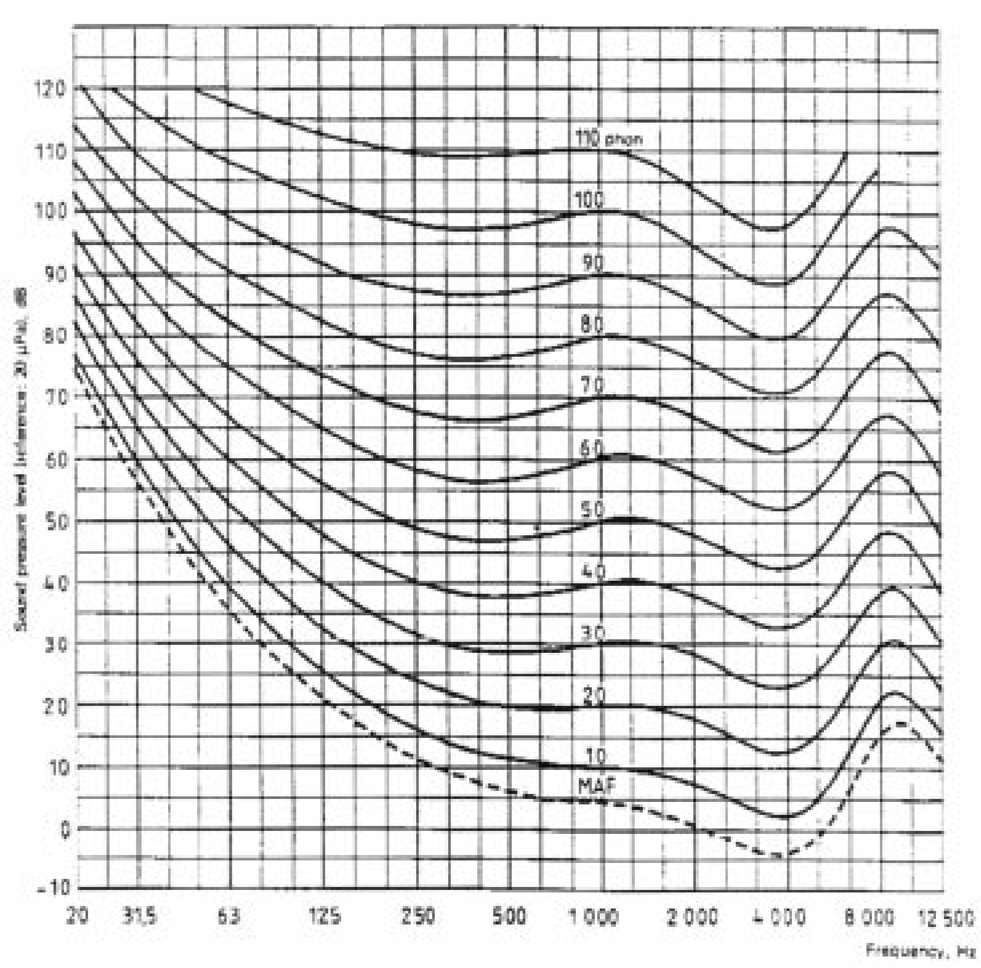

Figure 1: Isophones [Roe79]

Therefore, the loudness levels of (f_s, L_1) and (f_l, L_3) are identical (say: $L_{N,1}$) as are those of (f_s, L_2) and (f_l, L_4) (say: $L_{N,2}$). Next, we calculate the relative position of (f_x, L_x) between the two straight lines at $f = f_x$ and calculate the same relative position between $L_{N,1}$ and $L_{N,2}$ at $f = 1\ kHz$, which gives us the required loudness level of (f_x, L_x). In this last step, we assume the scaling of the loudness level to be the same at every frequency.

The loudness level relates tones of different frequency to each other. However, a tone whose loudness level is twice that of another tone does not sound twice as loud. Therefore, a relation between different loudness levels must be specified. This is performed by a measure called **loudness** N $[sone]$ [Roe79, ZF90]. The alterations in SPL necessary to double the perceived loudness can be seen in

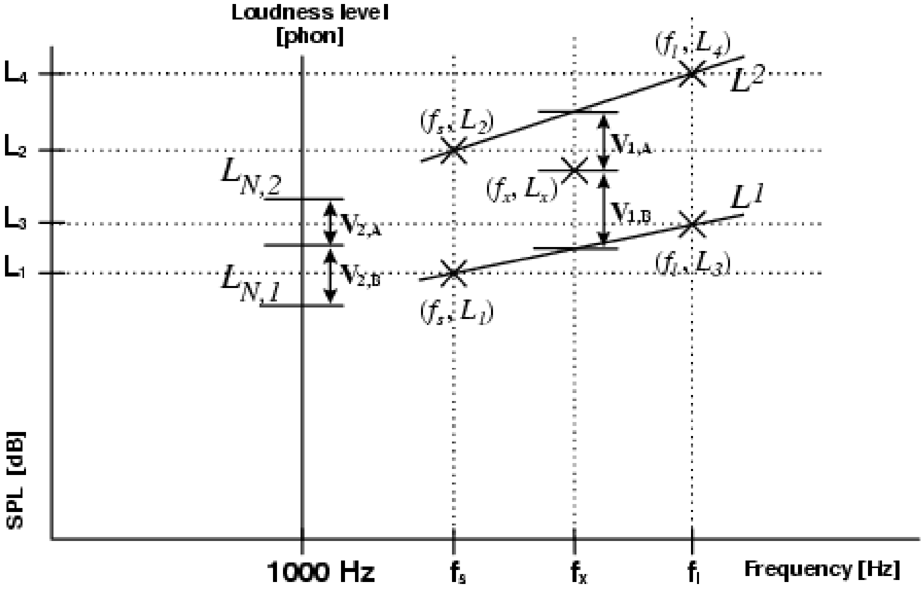

Figure 2: Algorithm to calculate approximation of loudness level

Figure 3, which may be described approximately by:

$$\Delta L = \begin{cases} 10dB & : \quad L \geq 40dB \\ \frac{1}{4}L & : \quad L < 40dB \end{cases} \tag{3}$$

From this, we can deduce a function for the calculation of the loudness. The loudness value for a loudness level of $L_N = 40\ dB$ is defined as 1 *sone*. For values above 40 dB, [Zwi82] gives an exponential function: $N = 2^{\frac{L-40\ dB}{10}}$ which is based on the above-mentioned SPL changes. As we did not find a function in the literature to describe loudness for $L_N < 40\ dB$, we deduce it from equation (3).

To double a given loudness value, we need to add $\frac{1}{4}L$. And we know that $N(40\ dB) = 1$ *sone*. We get the linear equation system:

$$2 \cdot N(L) = \quad N(\tfrac{5}{4}L) \tag{4}$$

$$N(40) = \quad 1 \tag{5}$$

Representing equation (4) logarithmically and substituting $M(\log_{10} L) = \log_{10} N(L)$ results in:

$$\log_{10} 2 + M(\log_{10} L) = \quad M(\log_{10} \tfrac{5}{4} + \log_{10} L) \tag{6}$$

In equation (6), $M(\log_{10} L)$ is a linear function, because when incrementing the value by $\log_{10} \frac{5}{4}$, the function value increases by $\log_{10} 2$:

$$M(\log_{10} L) = a \cdot \log_{10} L + b \tag{7}$$

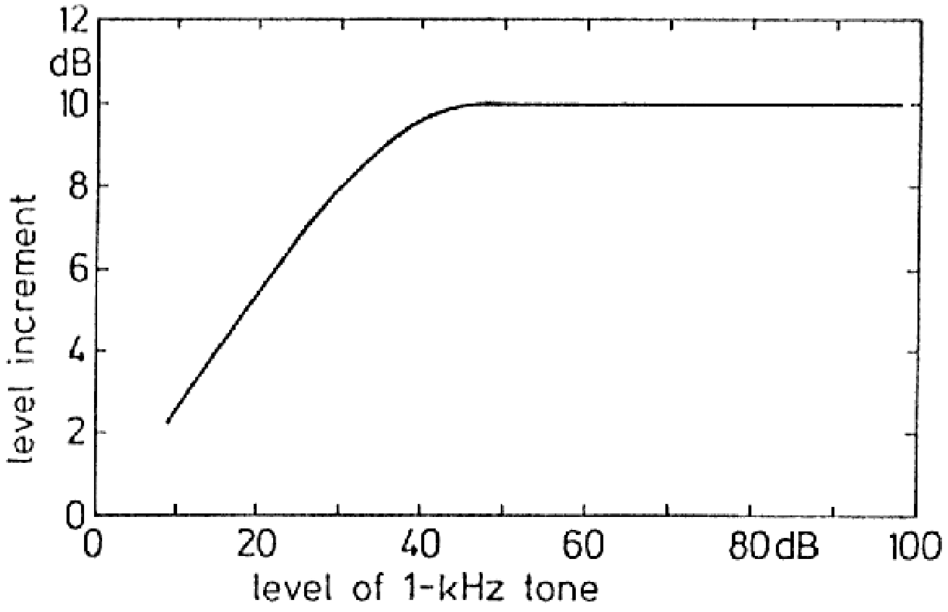

Figure 3: Necessary alterations in SPL to double loudness [ZF90]

From equation (6), we can deduce the gradient a:

$$a = \frac{\log_{10} 2}{\log_{10} \frac{5}{4}} \tag{8}$$

The value of b results from equation (5):

$$\begin{aligned} N(40) = & \quad 1 & \Rightarrow \\ \log_{10} N(40) = & \log_{10} 1 & \Rightarrow \\ \log_{10} N(40) = & \quad 0 \end{aligned}$$

together with equations (7) and (8):

$$\begin{aligned} 0 = & \quad a \cdot \log_{10} 40 + b & \Rightarrow \\ 0 = & \tfrac{\log_{10} 2}{\log_{10} \frac{5}{4}} \cdot \log_{10} 40 + b & \Rightarrow \\ b = & -\tfrac{\log_{10} 2}{\log_{10} \frac{5}{4}} \cdot \log_{10} 40 \end{aligned} \tag{9}$$

Finally, when inserting equations (8) and (9) into equation (7), and substituting back $\log_{10} N(L) = M(\log_{10} L)$, we arrive at the loudness function for $L_N < 40\ dB$:

$$M(\log_{10} L) = \tfrac{\log_{10} 2}{\log_{10} \frac{5}{4}}(\log_{10} L - \log_{10} 40) \quad \Rightarrow$$

$$
\begin{aligned}
M(\log_{10} L) = & \quad \log_{10}(\tfrac{L}{40}) \cdot \tfrac{\log_{10} 2}{\log_{10} \frac{5}{4}} & \Rightarrow \\
\log_{10} N(L) = & \quad \log_{10}(\tfrac{L}{40}) \cdot \tfrac{\log_{10} 2}{\log_{10} \frac{5}{4}} & \Rightarrow \\
N(L) = & \quad (\tfrac{L}{40})^{\frac{\log_{10} 2}{\log_{10} \frac{5}{4}}} &
\end{aligned}
$$

A closed form for the loudness function results when the SPL L is replaced by equation (1):

$$
N = \begin{cases} 2^{(\log_{10}(\frac{I_{1kHz}}{I_0})-4)} & : \quad L = 10 \ \log_{10} \frac{I_{1kHz}}{I_0} \geq 40 \ dB \\ (\frac{\log_{10} \frac{I_{1kHz}}{I_0}}{4})^{\frac{\log_{10} 2}{\log_{10} \frac{5}{4}}} & : \quad L = 10 \ \log_{10} \frac{I_{1kHz}}{I_0} < 40 \ dB \end{cases} \tag{10}
$$

This approximated loudness function can be seen in Figure 4. It only relates to the 1 kHz tone, but the abscissa may also be labeled with the loudness level L_N such that this function explains loudness differences of two different sinus tones A and B.

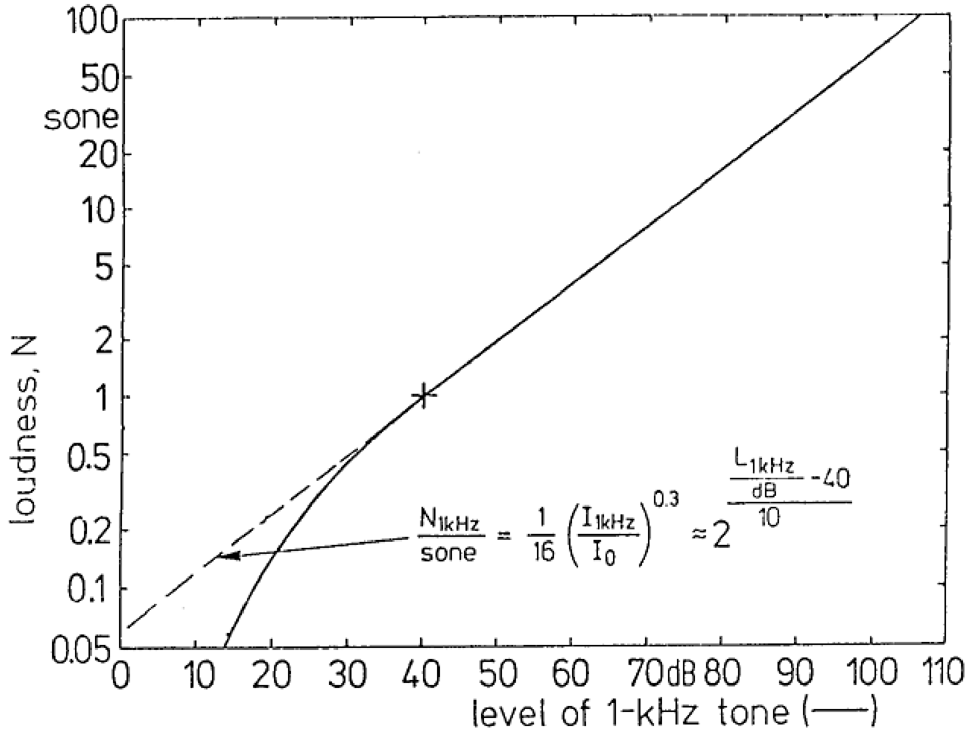

Figure 4: Function of loudness [ZF90]

In order to determine loudness judgements of complex tones, we need to integrate loudness measures of different sinus tones. Psychoacoustical experiments, however, have found out that it does not suffice to simply add up individual loudness measures [Zwi82]. The integral loudness of two sinus tones which are close in frequency, is calculated by adding their SPLs. If, however, they are far enough apart, their loudness may be simply added. This critical distance is described by so-called **frequency groups**. It is dependent on the frequency and can be seen in Figure 5.

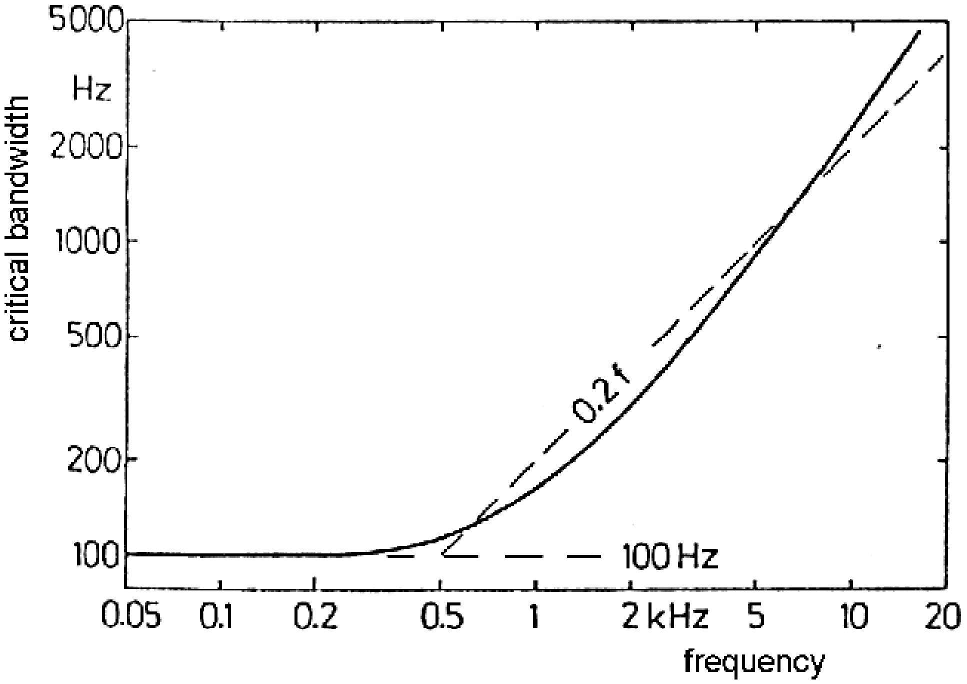

Figure 5: Width of frequency groups [Hel93]

For tonality measures, a table was standardized for lined up frequency groups (see table 1). We use this table to determine whether two tones are within the same frequency group. This is not quite the same as what the ear does, because the ear does not divide absolute frequency groups, but categorizes dependent upon the appearing frequencies. However, our approach gives a good approximation.

Calculation of the **integral loudness** of a complex tone is now performed as follows:

1. Calculate the frequencies (and their SPL) of the partial tones in the sound (we used the Fast Fourier Transform with a window size of 1024 samples, an overlap of 142 samples and a Hamming windowing function).

2. Sum up the SPLs within each frequency group.

3. Calculate the loudness of each frequency group.

4. Sum up the loudnesses of all frequency groups.

z [Bark]	f_u, f_o [Hertz]	Δf [Hertz]	z [Bark]	f_u, f_o [Hertz]	Δf [Hertz]
1	0, 100	100	13	1720, 2000	280
2	100, 200	100	14	2000, 2320	320
3	200, 300	100	15	2320, 2700	380
4	300, 400	100	16	2700, 3150	450
5	400, 510	110	17	3150, 3700	550
6	510, 630	120	18	3700, 4400	700
7	630, 770	140	19	4400, 5300	900
8	770, 920	150	20	5300, 6400	1100
9	920, 1080	160	21	6400, 7700	1300
10	1080, 1270	190	22	7700, 9500	1800
11	1270, 1480	210	23	9500, 12000	2500
12	1480, 1720	240	24	12000, 15500	3500

Table 1: Tonality and frequency groups [Zwi82]

This integral loudness algorithm represents a means to predict from the sound pressure levels of a sound file the level of loudness as perceived by a human. We are aware that there are more influences on the human perception of loudness than the ones we included. For example, the duration of a tone also influences the loudness perceived.

3 Experiments

As this work is part of the MoCA project, we used mainly audio material from feature films and TV commercials for our experiments. In addition, we used a piece of classical music. We have extracted representative material of 30 min duration (digitization details: 22050 Hz sampling rate, mono, 8 bit precision). This was played on a computer to test people. In order to make it feasible for the humans to give judgements, we intoduced a loudness scale consisting of five different classes and called them *pp*, *p*, mf, f and ff analogous to the notation of dynamics in music.

The results of the human loudness judgements were compared to those of the computer. Therefore, we divided the calculated integral loudness (IL) values and also the sound pressure level (SPL) values into the same five classes. We first let the humans give a judgement of the loudest part to calibrate the program results. Next, we solved the question of how to divide up the five classes. Several experiments suggested that a linear partitioning does not work as well as an arithmetic partitioning. The arithmetic partitioning which we use for the IL values works as follows: if S is the size of *pp*, then $2S$ is the size of *p*, $3S$ that of mf, etc. The arithmetic partitioning used for the SPL values, however, had to be done the other way round in order to get any valuable results from the SPL:

Number	Description	Type
1	60 sec commercial, 60 sec Emergency Room	commercial & series
2	60 sec Forest Gump, 60 sec commercial	feature film & commercial
3	60 sec Melrose Place, 60 sec commercial	series & commercial
4	60 sec commercial, 60 sec Interview with a vampire	commercial & feature film
5	300 sec Platoon (war movie)	feature film
6	300 sec Sea of Love (thriller)	feature film
7	648 sec Allegro in G-Major, Mozart	classic music

Table 2: Material used for experiments

if S is the size of ff, then $2S$ is the size of f, $3S$ that of mf, etc.

Table 2 gives an overview of the material used in the experiments. In order to avoid a disturbing cut, the material was faded in and out over one second. The calibration produced a loudness perception of ff for each piece except for piece 3, which received a f for its loudest part.

The five test people chosen were all between 18 and 30 years old in order to reduce possible effects of reduced hearing capabilities on loudness judgements. The tests were all performed using the same sound system and the same loudspeakers in a relaxed atmosphere so as to minimize pressure. But people still needed 5 to 10 min of practice before they became accustomed to the task. They also had difficulties concentrating on the task, especially during the long piece No 7. One person even stated that after a while she judged her perception of pitch rather than the loudness of this piece! We also discovered that the sensitivity tochanges in loudness is not equal in all people - some changed their judgement more often while others equalized more.

Figures 6 and 7 show as an example the results of the IL and the SPL measures compared to the human loudness judgements for the test pieces 6 and 7, accumulated to seconds. These give a rough idea of how well the IL represents the human perception of loudness.

A detailed analysis of the similarity was performed on all test pieces. We calculated the average deviation (on the scale of 1="pp" to 5="ff"), variance and correlation of the integral loudness and the SPL from the average human loudness judgement. The results can be seen in table 3. Interesting is a comparison of these results with the average deviation, variance and correlation of the results of the five test people with their average loudness judgement (see table 4).

Clearly, the IL measure always surpasses the SPL measure, with as much as 10 to 20% less (average) deviation. The variance of the deviation and correlation

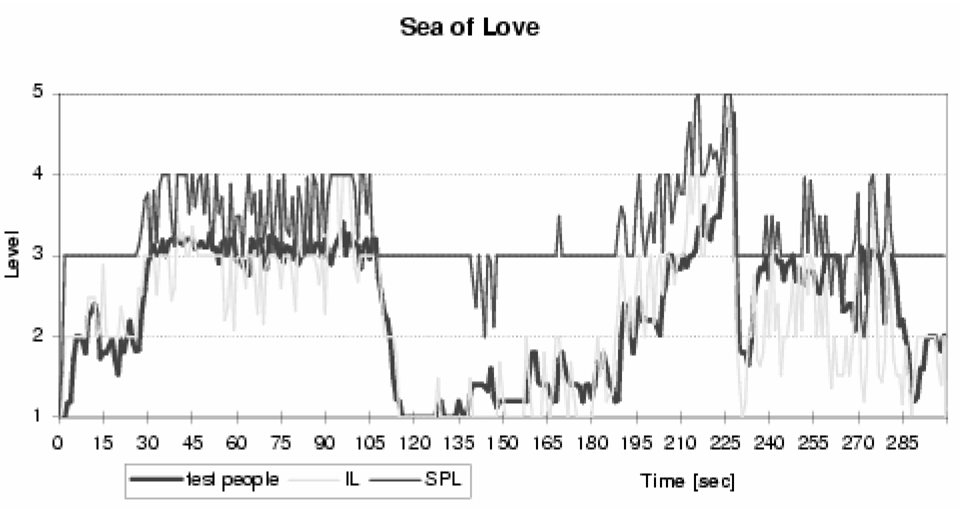

Figure 6: Comparison of IL and SPL with human perception of loudness of piece 6

are also unmistakably better. Yet, the IL results still are not nearly as good as the results of the humans. A look at the diagrams (see Figures 6 and 7) shows that this might be due to a considerable delayed human reaction. A human cannot press the button as fast as a computer can react to a change in loudness. To prove this, we delayed the computer results of piece 3 by one second. This resulted in an increase in correlation by 10% for both SPL and IL. An additional delay by one second failed to increase these results further.

4 Conclusion and Outlook

The experiments with the two loudness measures SPL and IL showed that it makes a significant difference whether or not a loudness measure is modelled according to human perception. Our suggested measure of cognitive loudness IL already very closely approximates the human measure of loudness (see e.g. the average results of person P4 compared to the average results of IL). The mathematical operations necessary to perceptualizing the SPL are expensive, however, especially when applied during a realtime feature extraction process from audio. Therefore, more detailed experiments on the value of each of the proposed processing steps is necessary. It could well be that not all steps need to be performed to achieve a significant improvement.

Overall, we can say that if perceptive adaptation is already important with reagrd to loudness measurements, it must be even more important for more complex sound measures like pitch and timbre. These questions remain to be answered in the future.

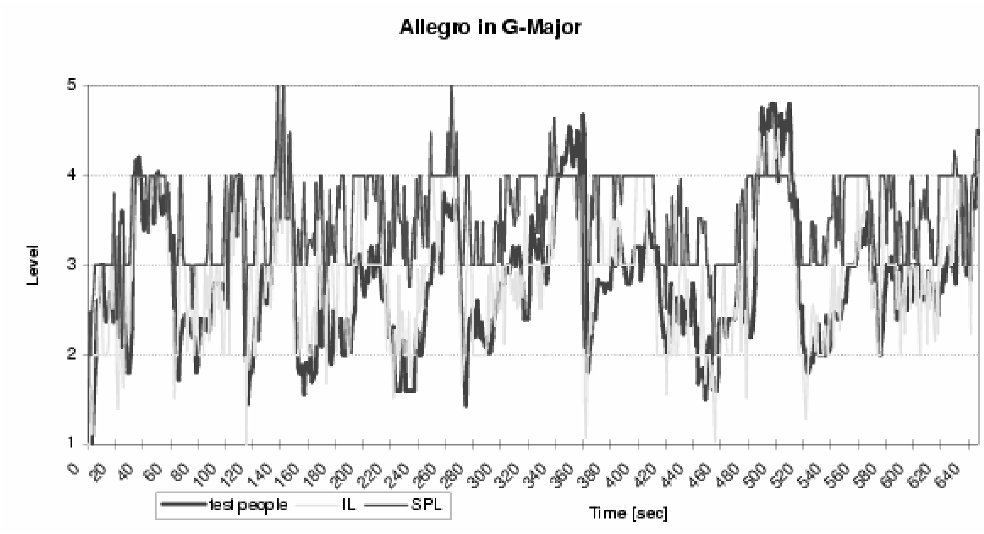

Figure 7: Comparison of IL and SPL with human perception of loudness of piece 7

Acknowledgement

Thanks to Karl Rupilius who performed most of the theoretical and programming work for this paper in the course of his diploma work.

Number	SPL/IL	avg. deviation from human perception	variance of avg. deviation	correlation with human perception
1	SPL	0.60	0.25	69 %
	IL	0.58	0.22	73 %
2	SPL	0.71	0.26	87 %
	IL	0.53	0.12	91 %
3	SPL	1.37	0.27	69 %
	IL	0.42	0.17	76 %
4	SPL	0.53	0.17	85 %
	IL	0.48	0.12	87 %
5	SPL	0.70	0.26	75 %
	IL	0.60	0.28	79 %
6	SPL	0.99	0.35	64 %
	IL	0.42	0.12	84 %
7	SPL	0.81	0.22	58 %
	IL	0.43	0.12	73 %
average	SPL	0.82	0.25	72 %
	IL	0.49	0.16	80 %

Table 3: Deviation factors between loudness measures and human loudness perception

Number	Person	avg. deviation from avg. human perc.	variance of avg. deviation	correlation with avg. human perc.
1	P1	0.35	0.09	93 %
	P2	0.32	0.09	91 %
	P3	0.35	0.12	83 %
	P4	0.38	0.11	86 %
	P5	0.33	0.11	78 %
2	P1	0.35	0.04	97 %
	P2	0.37	0.05	96 %
	P3	0.37	0.05	87 %
	P4	0.58	0.17	90 %
	P5	0.42	0.08	96 %
3	P1	0.34	0.10	85 %
	P2	0.29	0.06	88 %
	P3	0.33	0.84	86 %
	P4	0.51	0.22	76 %
	P5	0.31	0.09	92 %
4	P1	0.41	0.07	94 %
	P2	0.32	0.04	95 %
	P3	0.42	0.09	88 %
	P4	0.36	0.06	89 %
	P5	0.39	0.05	94 %
5	P1	0.49	0.19	86 %
	P2	0.40	0.12	94 %
	P3	0.38	0.12	81 %
	P4	0.50	0.15	82 %
	P5	0.39	0.13	89 %
6	P1	0.24	0.05	93 %
	P2	0.26	0.07	93 %
	P3	0.27	0.07	82 %
	P4	0.53	0.17	77 %
	P5	0.32	0.11	91 %
7	P1	0.44	0.12	81 %
	P2	0.46	0.11	83 %
	P3	0.55	0.16	86 %
	P4	0.67	0.24	69 %
	P5	0.46	0.14	65 %
average	P1	0.37	0.09	90 %
	P2	0.35	0.08	91 %
	P3	0.38	0.21	85 %
	P4	0.50	0.16	81 %
	P5	0.37	0.10	86 %

Table 4: Deviation factors between single persons and average human loudness perception

References

[Hel93] J. Hellbrück. *Hören - Physiologie, Psychologie und Pathologie.* Hogrefe, Verlag für Psychologie, Göttingen, Bern, Toronto, Seattle, 1993. (in German).

[ISO61] ISO, International Organization for Standardization. Recommendation ISO R 226, Normal-loudness contours for pure tones and normal threshold of hearing under free field listening conditions, 1961.

[ISO87] ISO, International Organization for Standardization. International standard 226, Acoustics - normal equal-loudness level contours, 1987.

[Mol73] J. A. Molino. Pure-tone equal-loudness contours for standard tones of different frequencies. *Percept. Psychophys.*, 14(1), 1973.

[Roe79] J.G. Roederer. *Introduction to the physics and psychophysics of music.* Springer, New York, 1979.

[ZF90] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models.* Number 22 in Springer Series in Information Sciences. Springer Verlag, Berlin, Heidelberg, 1990.

[ZFS57] E. Zwicker, G. Flottorp, and S. S. Stevens. Critical bandwidth in loudness summation. *J. Acoustical Society of America*, 29:548 ff., 1957.

[Zwi82] E. Zwicker. *Psychoakustik.* Springer Verlag, Berlin, Heidelberg, 1982. (in German).