Proceedings of Meetings on Acoustics

Volume 28 http://acousticalsociety.org/

22nd International Congress on Acoustics
Acoustics for the 21st Century

Buenos Aires, Argentina
05-09 September 2016

Psychological and Physiological Acoustics: Paper ICA2016 - 113

Modelling the sensation of fluctuation strength

Alejandro Osses Vecchi and **Rodrigo García León**
Human-Technology Interaction Group, Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, the Netherlands; a.osses@tue.nl; yo@rodrigogarcia.me

Armin Kohlrausch
Human-Technology Interaction Group, Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, the Netherlands; Brain Behaviour & Cognition Group, Philips Research Europe, Eindhoven, the Netherlands; a.kohlrausch@tue.nl

The sensation of fluctuation strength (FS) is elicited by slow modulations of a sound, either in amplitude or frequency (typically <20 Hz), and is related to the perception of rhythm. In speech, such periodicities convey valuable information for intelligibility (prosody). In western music, most of the envelope periodicities are also found in that range. These are evidences of the potential relevance of FS in the perception of speech and music. There is, however, no published computational model to assess the FS of a sound. This might be one reason why when slow modulations of a sound are to be analysed, other indirect measures (e.g., loudness to estimate "loudness fluctuations") or more complex techniques (e.g., the modulation filter bank) are used. In this paper a model of fluctuation strength is presented. The model was developed taking advantage of the physical similarity between FS and the psychoacoustical sensation of roughness. The FS model was then adjusted and fitted to existing experimental data collected using artificial stimuli, namely, amplitude- (AM) and frequency- (FM) modulated tones and amplitude-modulated broadband noise (AM BBN). The test battery of sounds also considered samples of male and female speech and some musical instrument sounds.

Published by the Acoustical Society of America

1. INTRODUCTION

Temporal fluctuations in amplitude and in frequency are found naturally in everyday sounds. Amplitude modulations (AM) are related to the envelope of a waveform, while frequency modulations (FM) to its fine structure. Envelope refers to the perceived acoustic amplitude of a sound that is integrated by the hearing system due to its slow response (or “sluggishness”) to high rate (sound pressure) variations of its waveform. Two examples of everyday sounds are speech and music. Speech was described by Rosen [1] as temporal fluctuating patterns with three partitions: envelope, periodicity and fine structure. The envelope contributes to, among other factors, prosody (i.e., duration, speech rhythm) and articulation, periodicity to intonation and fine structure to the timbre of a sound. With these concepts, it seems logical to assume that the characterisation of speech as a temporal fluctuating pattern is also applicable to music. The link between prosody and Western music found by Patel et al. [2] supports this assumption.

Two of the well-known classical psychoacoustical metrics are related to the perception of modulated sounds: fluctuation strength (FS) [3, 4] and roughness [5], for sounds modulated at slower frequencies (<20 Hz) and more rapid modulation rates (20-300 Hz), respectively. Both sensations show a bandpass characteristic with peaks at 4 Hz for FS and 70 Hz for roughness. The range of modulations below 20 Hz has been shown to be of special interest for speech intelligibility [6, 7] as well as for the perception of rhythm, which is related to the average syllable rate at AMs of around 4 Hz [8].

FS is an attribute related to the perception of the envelope in the range that we indicated as relevant for speech intelligibility (and potentially also for music). Roughness, however, is an attribute related to timbre (due to the higher modulation frequency range) that has taken more attention for its accepted influence in the perception of unpleasantness of a sound. There are, therefore, a number of published roughness models [e.g., 5, 9, 10]. Less detailed information about the algorithms to assess FS is available or solutions that apply for a specific type of stimuli have been described, for instance for AM sinusoids or AM broad-band noise (AM BBN) [3, 11]. Examples of the first case are the algorithms available in commercial software packages (Pulse by Brüel & Kjær, ArtemiS by Head Acoustics GmbH, PAK by Müller BBM, PAAS [12]). In this paper a model of FS is presented. The similarities between FS and roughness listed above motivated the development of our implementation based on an existing roughness model [9, 13]. Although a similar approach was followed by Sontacchi almost 20 years ago [12] our database of sounds used for developing and testing the algorithm is more diverse, including not only artificial sounds (AM and FM tones and AM BBN) but also a few cases of male and female speech and music samples, which were taken from the test battery of sounds used in [14]. One of the goals of this paper was to give the first steps towards the development of a “unified” FS model in line with previous research quantifying how close our results are from estimates provided in the literature, obtained either experimentally or by using other computational algorithms.

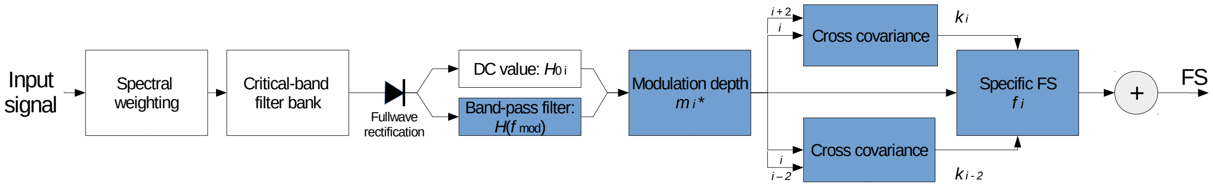

Figure 1. Structure of our model of fluctuation strength. The highlighted blocks represent the processing stages that were modified with respect to the reference roughness models.

2. METHODS

A. Model of fluctuation strength

The algorithm used in the FS model was adapted from the roughness extraction algorithm described in [5, 9]. The structure of the model is shown in Figure 1. The model assumes that the total FS is the sum of partial contributions from N auditory filters and it is based on the concept of modulation:

$$FS = \sum_{i=1}^{N} f_i = C_{FS} \cdot \sum_{i=1}^{N} (m_i^*)^{p_m} \cdot |k_{i-2} \cdot k_i|^{p_k} \cdot \left(g(z_i)\right)^{p_g} \tag{1}$$

where N is the number of auditory filters (here N=47), m^* is a generalised modulation depth, k refers to the normalised cross covariance between different auditory filters and $g(z_i)$ is an additional free parameter to introduce a weighting as a function of centre frequency. The product of all the elements in Eq. (1) as a function of the critical band i defines the specific fluctuation strength f_i. The parameters C_{FS}, p_m, $p_{\underline{k}}$ and p_g are constants optimised to fit the model. Further explanation of these parameters is provided in the subsequent sections.

In general, the model provides FS estimates for successive analysis frames. The frames have a duration of 2 s and a 90%-overlap and are gated on and off with 50-ms raised-cosine ramps.

Each analysis frame is independently and successively passed through the processing blocks described below. For this reason from hereafter we refer to all analysis frames as the "input signal".

i. Spectral weighting

To approximate the incoming signal to what arrives to the oval window (beginning of the inner ear), a transmission factor a_0 is applied. This factor introduces a frequency dependent gain that accounts for the sound transmission from free-field through the outer and middle ear. In the model a_0 was implemented as a 4096th-order FIR filter.

ii. Critical-band filter bank

In the frequency domain (N-point FFT, frequency resolution $\Delta f = 0.5$ Hz), all frequency bins with amplitudes above the absolute hearing threshold are transformed into a triangular excitation pattern [15]. The triangular excitation pattern produced by the frequency component f (in Hz) at a level L (in dB) has a constant lower slope S_1 of 27 dB/Bark and higher slope S_2 defined by Eq. (2).

$$S_2 = 24 + \frac{230}{f} - 0.2 \cdot L \tag{2}$$

The slopes S_1 and S_2 are defined in the frequency domain and referred to the critical-band scale, expressed in Bark. An analytical expression to relate the frequencies z in Bark and f in Hz is given by Eq. (3) [16].

$$z = 13 \cdot arctan(0.76 \cdot 10^{-4}\, f) + 3.5 \cdot arctan\left(\left[\frac{f}{7500}\right]^2\right) \tag{3}$$

The excitation patterns are a way to determine the contribution of a given frequency f_k (and level L_k) to another auditory filter, located at an "observation point" i, with a Bark distance of Δz Bark (keeping the same phase of the component at k). That contribution, $L_{k,i}$, can be expressed as:

$$L_{k,i} = \begin{cases} L_k - S_2 \cdot \Delta z = L_k - S_2 \cdot (z_i - z_k) & if\ f_k < f_i \\ L_k - S_1 \cdot \Delta z = L_k - S_1 \cdot (z_k - z_i) & if\ f_k > f_i \end{cases} \tag{4}$$

where z_i and z_k are the corresponding frequencies f_i and f_k in the critical-band rate scale that can be calculated using Eq. (3).

If we now consider 47 equally spaced "observation points" (with a spacing of 0.5 Bark) related to the frequency range from 0.5 Bark (50 Hz) to 23.5 Bark (13.2 kHz) and evaluate the individual contribution of each computed excitation pattern, 47 output (audio) signals are obtained. These 47 signals can be interpreted as the output of a critical-band filter bank with centre frequencies $z_i = (0.5 \cdot i)$ Bark and bandwidth of 1 Bark. At the end of this stage, each spectrum is converted back to the time domain using an inverse Fourier Transform (IFFT), obtaining 47 $e_i(t)$ signals.

iii. Generalised modulation depth m^*

Each of the 47 signals $e_i(t)$ obtained from the critical filter bank is used to obtain an estimate of the modulation depth m^*. The so-called generalised modulation depth is calculated by dividing the root-mean-square value (RMS) of the weighted envelopes of $h_{BPi}(t)$ by their DC values $h_{0,i}$. The DC value is calculated from the full-wave rectified time signals:

$$h_{0,i} = \left|\overline{e_\iota(t)}\right| \tag{5}$$

The weighted excitation envelopes are obtained by filtering each full-wave rectified signal $|e_i(t)|$ using a bandpass filter $H(f_{mod})$. This filter is applied in the envelope domain, introducing a weighting in the lower part of the excitation patterns. The shape of the $H(f_{mod})$ function was chosen to account for the bandpass characteristic of the sensation of fluctuation strength (with maximum at a modulation frequency of 4 Hz). The resulting $H(f_{mod})$ was implemented as an IIR filter with passband between 3.1 and 12 Hz (see section 3.A for further details).

The RMS of the weighted functions $\overline{h_{BP,\iota}}$ is then used to obtain the generalised modulation depths:

$$m_i^* = \frac{\overline{h_{BP,\iota}}}{h_{0,i}} \tag{6}$$

In the original roughness model this ratio was limited to a maximum value of 1. FM tones represent a case where this limitation was often being applied, but their roughness in asper reaches larger values (3.2 asper for a 1.6-kHz tone, f_{mod} at 80 Hz, f_{dev} of ±800 Hz and 60 dB SPL) than those for FS in vacil (1.4-kHz tone, f_{mod} at 4 Hz, f_{dev} of ±700 Hz and 60 dB SPL). In the FS model we suggest the introduction of a compression stage to the ratio m_i^* rather than a limitation. A compression ratio of 3:1 is applied when the modulation depth estimate exceeds a threshold of 0.7 units. This means that if m_i^* is 0.15 units above the threshold, i.e., $m_i^*{}_{input}$ =0.85, the resulting modulation depth will be 0.05 (0.15/3) above threshold resulting in $m_i^*{}_{output}$ =0.75.

iv. Normalised cross covariance

In a discrete time domain the normalised cross covariance (in short, cross covariance) between the functions *x* and *y*, both being *N* samples long, is defined by Eq. (7) [see e.g. 17, their Eq. (2)]:

$$k = \frac{\sum x \cdot y - \frac{1}{N} \cdot \sum x \cdot \sum y}{\sqrt{\left[x^2 - \frac{1}{N} \cdot (\sum x)^2\right] \cdot \left[y^2 - \frac{1}{N} \cdot (y)^2\right]}} \tag{7}$$

Within the FS model the cross covariance between adjacent critical bands is assessed to determine whether their modulations are in or out of phase. The more in-phase the modulations are determines to what extent the specific FS can be summed up to obtain the total FS. In this manner, the cross covariance between the channel *i* and the channels one Bark below (i–2) and above (*i*+2) are computed. In other words, to obtain the factor $k_{i\text{-}2}$, *x* and *y* in Eq. (7) have to be replaced by $h_{BP,i\text{-}2}$ and $h_{BP,i}$, respectively. Likewise, to obtain the factor k_i, *x* and *y* have to be replaced by $h_{BP,i}$ and $h_{BP,i+2}$.

B. Stimuli

In order to fit and validate the FS model we chose a set of stimuli with known FS values. Part of the set corresponded to artificial stimuli: AM tones, FM tones and AM BBN. The rest of the test stimuli were chosen from everyday sounds. The reference sound to which an FS of 1 vacil is ascribed is an AM sine tone centred at f_c=1000 Hz, modulated at an $\mathrm{f}_{\mathrm{mod}}$ of 4 Hz and level of 60 dB. A summary of the artificial stimuli used in the validation is shown in Table 1. For this set of stimuli, FS values obtained in perceptual experiments are available from the literature [11]. Additionally, a set of everyday stimuli, particularly speech and music samples, were chosen from the database of sounds used in [14]. That database consists of 70 sounds, out of which 7 representative sound samples were chosen. The selection of the samples was as follows: (a) three representative speech samples (one male voice, one female voice, babble noise); (b) two music samples of soloist and ensemble playing, and (c) the sounds having minimum and maximum FS. For that database, Schlittmeier *et al*. [14] used a commercial software to obtain their FS values. The selected samples are summarised in Table 2.

Table 1. Artificial stimuli used to validate the FS model. FS values from the literature [11] ***are also shown.***

Type	Fixed parameters	SPL [dB]	Variable parameters (FS)
AM tone (reference)	f_c=1000 Hz m_{index}=1	60	f_{mod} ={4.00} Hz (1.00) vacil
AM tone	f_c=1000 Hz m_{index}=1	70	f_{mod} ={1.00, 2.00, 4.00, 8.00, 16.0, 32.0} Hz (0.39, 0.84, 1.25, 1.30, 0.36, 0.06) vacil
FM tone	f_c=1500 Hz f_{dev}=±700 Hz	70	f_{mod} ={1.00, 2.00, 4.00, 8.00, 16.0, 32.0} Hz (0.85, 1.17, 2.00, 0.70, 0.27, 0.02) vacil
AM BBN	BW=16000 Hz m_{index}=1	60	f_{mod} ={1.00, 2.00, 4.00, 8.00, 16.0, 32.0} Hz (1.12, 1.58, 1.80, 1.57, 0.48, 0.14) vacil

Table 2. Everyday sounds used to validate the FS model. An artificial noise (pink noise, Track Nr. 61) was also included. The average sound pressure level (SPL) of each sound sample is shown. For the changing-state speech samples and the ducks' quacking samples the maximum levels are also shown. The FS values were taken from [14] ***and they were computed using a commercially available algorithm***

Type	**Track Nr. / Description**	**SPL [dB]** L_{eq} (L_{max})	**Reported FS [vacil]**
Speech	1 / Narration, female voice	56.1 (67.2)	1.11
Speech	2 / Narration, male voice	60.0 (69.4)	1.21
Speech	23 / Eight talker babble noise	63.6 (67.8)	0.38
Music	24 / Strings concert	62.1	0.21
Music	31 / Violin solo	58.2	0.56
Animal	34 / Ducks' quacking	64.5 (73.4)	1.77
Noise*	61 / Broadband (pink) continuous noise	60.1	0.02

3. RESULTS

A. Artificial stimuli

The artificial stimuli were used to fit the free parameters of the model: the constant C_{FS}, the bandpass filter $H(f_{\mathrm{mod}})$ and the exponents p_m and p_k. First, the reference sound, that has a fluctuation strength of 1 vacil, was used to set the constant C_{FS}. A value C_{FS} of 0.2490 was found. Subsequently, the bandpass filter $H(f_{\mathrm{mod}})$ was fitted by using 1-kHz AM tones with f_{mod} from 1 to 32 Hz with the exponents p_m=p_k=1.7 and p_g=1 ($g(\mathrm{z_i})$ was initially set to 1 for all i values, i.e., no weighting is considered). As a result two cascade IIR filters (4th-order LPF and 2nd-order HPF) producing a bandpass filter between 3.1 and 12 Hz were obtained. As can be seen in Figure 2, so far the fitted model predicts qualitatively the fluctuation strength for AM tones, FM tones and AM BBN, although the FS for the FM tones is overestimated for modulation frequencies above 4 Hz. Finally, some fine adjustments were introduced by reducing $g(z_i)$ gradually from 1 to 0.5, starting with the band centred at z_i=15 Bark (2.7 kHz) up to the band centred at z_i=23.5 Bark (13.2 kHz).

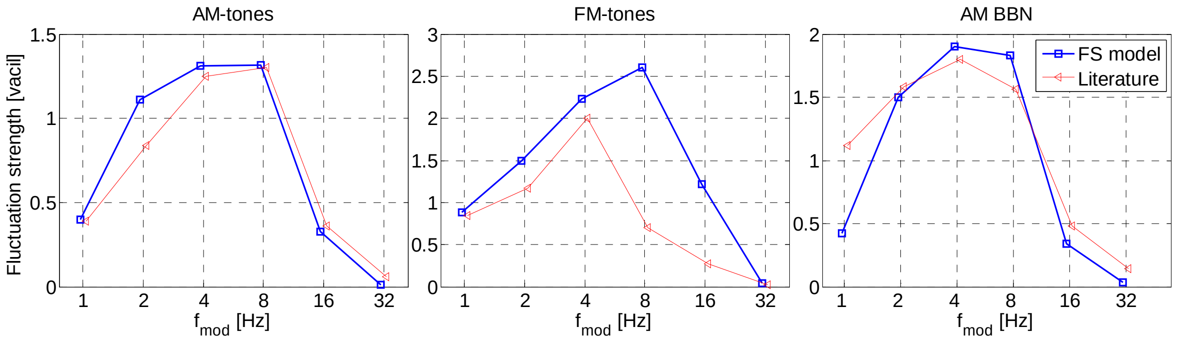

Figure 2. Results obtained from the FS model for: (left panel) AM tones; (middle panel) FM tones and (right panel) AM Broad-band noise.

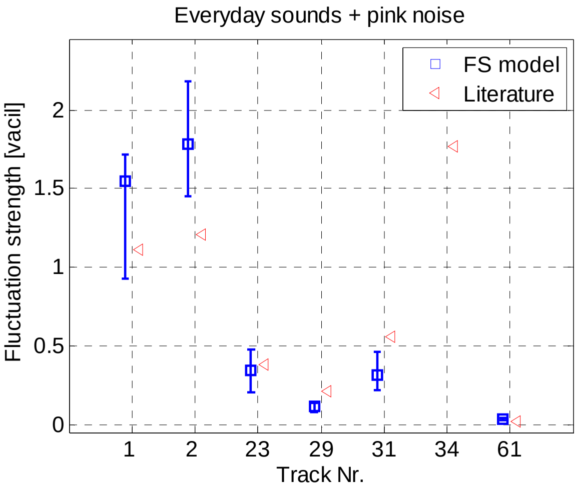

Figure 3. Results obtained from the FS model using the everyday sounds detailed in Table 2. The FS shown in squared markers correspond to median values along the sample duration. The errorbars represent the minimum and maximum FS. An extremely high FS value (4.2 vacil) was found for track 34 (Ducks' quacking, not shown in the figure).

B. Everyday sounds

The FS values given by the model for the everyday sounds (and pink noise) of Table 2 are shown in Figure 3. For the speech samples (Tracks 1 and 2) the median FS values were higher than the reference values by 0.45 and 0.58 vacil. For the eight-talker babble noise (Track 23), string concert (Track 29) and the pink noise (Track 61), the FS estimates seem to be in line with the reference values. For the violin solo (Track 31) there is an underestimation of the FS (difference of 0.25). The highest FS estimate was found for the ducks' quacking (FS of 4.2 vacil). This value was omitted in the figure since it is an "unreasonable" high estimate.

4. DISCUSSION AND CONCLUSION

As shown in the previous section, for a number of cases our FS model showed a reasonable agreement with FS estimates obtained either experimentally [4, 11] or by using commercially available software [14]. Particularly, within the subset of artificial stimuli there is a close agreement between the model and the experimental data for AM tones. Although the FS model shows a larger discrepancy for FM tones (overestimation) for modulation frequencies above f_{mod}=4 Hz, there is still a qualitative resemblance for the relation between FS and modulation rate. The maximum value of FS given by the model is shifted towards f_{mod}=8 Hz. Within the roughness model [see 9, their Fig. 9] a similar tendency was found, shifting the maximum roughness estimate to f_{mod}=80 Hz (instead of f_{mod}=70 Hz). Within the subset of everyday sounds, there is a good approximation between the FS values and the estimates reported in the reference paper for the eight-talker babble noise, the string concert and the pink noise samples. Although we found higher FS values for the male and female voices and the ducks' quacking sounds and a lower value for the violin sample, it is important to point out that the estimates presented in the reference paper were obtained from another FS algorithm and, therefore, it is unclear whether those FS values have been validated experimentally. Such an experimental validation for other sounds than those used in the original experimental work [4, 11] would be needed in order to evaluate the concept of FS and the various existing algorithms to compute it.

ACKNOWLEDGEMENTS

We would like to thank Sabine Schlittmeier for providing her database of everyday sounds. This research work has been funded by the European Commission within the ITN Marie Curie Action project BATWOMAN under the 7th Framework Programme (EC grant agreement Nr. 605867)

REFERENCES

[1] S. Rosen (1992). “Temporal information in speech: acoustic, auditory and linguistic aspects,” *Phil. Trans. R. Soc. London*, vol. 336. pp. 367–373.

[2] A. Patel, J. Iversen, and J. Rosenberg (2006). “Comparing the rhythm and melody of speech and music: the case of British English and French.,” *J. Acoust. Soc. Am.*, vol. 119, pp. 3034–3047.

[3] H. Fastl (1982). “Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise,” *Hear. Res.*, vol. 8, pp. 59–69.

[4] H. Fastl (1983). “Fluctuation strength of modulated tones and broad-band noise,” in *Hearing - Physical bases and psychophysics*, pp. 282–288.

[5] W. Aures (1985). “Ein Berechnungsverfahren der Rauhigkeit,” *Acustica*, vol. 58, pp. 268–281.

[6] R. Drullman, J. Festen, and R. Plomp (1994). “Effect of temporal envelope smearing on speech reception,” *J. Acoust. Soc. Am.*, vol. 95, pp. 1053–1064.

[7] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid (1995). “Speech recognition with primarily temporal cues,” *Science.*, vol. 270, pp. 303–304.

[8] V. Leong, M. a Stone, R. Turner, and U. Goswami (2014). “A role for amplitude modulation phase relationships in speech rhythm perception,” *J. Acoust. Soc. Am.*, vol. 136, pp. 366–381.

[9] P. Daniel and R. Weber (1997). “Psychoacoustical roughness: implementation of an optimized model,” *Acust. - Acta Acust.*, vol. 83, pp. 113–123.

[10] A. Kohlrausch, D. Hermes, and R. Duisters (2005). “Modeling roughness perception for sounds with ramped and damped temporal envelopes,” *Forum Acusticum*, pp. 1719–1724.

[11] H. Fastl and E. Zwicker (2007). “Fluctuation strength,” in *Psychoacoustics, Facts and Models*, 3rd edition, Springer Berlin Heidelberg, pp. 247–256.

[12] A. Sontacchi (1998). “Entwicklung eines Modulkonzeptes für die psychoakustische Geräuschanalyse unter MATLAB,” Master thesis, Graz University of Technology.

[13] R. García (2015). “Modelling the sensation of fluctuation strength,” Master thesis, Eindhoven University of Technology.

[14] S. Schlittmeier, T. Weissgerber, S. Kerber, H. Fastl, and J. Hellbrück (2012). “Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength,” *Atten. Percept. Psychophys.*, vol. 74, pp. 194–203.

[15] E. Terhardt (1979). “Calculating virtual pitch,” *Hear. Res.*, vol. 1, pp. 155–182.

[16] E. Zwicker and E. Terhardt (1980). “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency,” *J. Acoust. Soc. Am.*, vol. 68, pp. 1523–1525.

[17] S. van de Par and A. Kohlrausch (1995). “Analytical expressions for the envelope correlation of narrow-band stimuli,” *J. Acoust. Soc. Am.*, vol. 98, pp. 3157–3169.