Using a deep neural network to speed up a model of loudness for time-varying sounds

JOSEF SCHLITTENLACHER,[1,*] RICHARD E. TURNER[2] AND BRIAN C.J. MOORE[1]

[1] *Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge, CB2 3EB, UK*

[2] *Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK*

The “time-varying loudness (TVL)” model calculates “instantaneous loudness” every 1 ms, and this is used to generate predictions of short-term loudness, the loudness of a short segment of sound such as a word in a sentence, and of long-term loudness, the loudness of a longer segment of sound, such as a whole sentence. The calculation of instantaneous loudness is computationally intensive and real-time implementation of the TVL model is difficult. To speed up the computation, a deep neural network (DNN) has been trained to predict instantaneous loudness using a large database of speech sounds and artificial sounds (tones alone and tones in white or pink noise), with the predictions of the TVL model as a reference (providing the "correct" answer, specifically the loudness level in phons). A multilayer perceptron with three hidden layers was found to be sufficient, with more complex DNN architecture not yielding higher accuracy. After training, the deviations between the predictions of the TVL model and the predictions of the DNN were typically less than 0.5 phons, even for types of sounds that were not used for training (music, rain, animal sounds, washing machine). The DNN calculates instantaneous loudness over 100 times more quickly than the TVL model.

INTRODUCTION

Glasberg and Moore (2002) described a model for predicting the loudness of time-varying sounds: the time-varying loudness (TVL) model. A block diagram of the model is shown in Figure 1. The model includes a sequence of stages to simulate the transmission of sound to the eardrum (Shaw and Vaillancourt, 1985), the transmission of sound through the middle ear (Aibara *et al.*, 2001), the frequency analysis that takes place in the cochlea (resulting in an auditory excitation pattern) (Glasberg and Moore, 1990), the creation of a specific loudness pattern (including the effects of the compression that occurs in the cochlea) (Moore and Oxenham, 1998), and summation of specific loudness across characteristic frequencies (CFs) (Zwicker and Scharf, 1965) to give instantaneous loudness. Within the model, frequency is transformed to the ERB$_N$-number scale, which has units Cams (Glasberg and Moore, 1990; Moore, 2012). This is a perceptually relevant scale comparable to a scale of distance along the basilar membrane. Instantaneous loudness is assumed to be an intervening

[*] Corresponding author: js2251@cam.ac.uk; currently at the Department of Neurosciences, University of Cambridge.

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. © The Authors. ISSN: 2596-5522.

variable, not available to conscious perception, although it has been shown that certain cortical regions show activity that is correlated with the instantaneous loudness calculated using the model (Thwaites *et al.*, 2016).

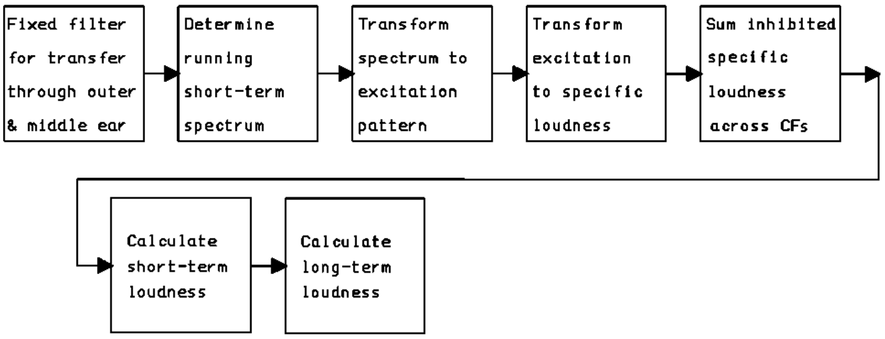
Fixed filter for transfer through outer & middle ear
Determine running short-term spectrum
Transform spectrum to excitation pattern
Transform excitation to specific loudness
Sum inhibited specific loudness across CFs
Calculate short-term loudness
Calculate long-term loudness

Fig. 1: Block diagram of the TVL model of Glasberg and Moore (2002).

The instantaneous loudness is smoothed over time to calculate short-term loudness, which is meant to represent the loudness of a short piece of sound such as a single word in a sentence or a note in a piece of music. The short-term loudness is itself further smoothed over time to calculate the long-term loudness, which is meant to represent the overall loudness of a longer stretch of sound, such as a whole sentence or a musical phrase. The peak value of the long-term loudness gives good predictions of judged loudness for a variety of sounds, including amplitude compressed speech (Moore *et al.*, 2003), sonic booms and impact sounds (Marshall and Davies, 2007), machinery sounds (Rennies *et al.*, 2015), and speech processed to have increased or decreased dynamic range (Zorila *et al.*, 2016). The model forms the basis of a proposed ISO standard (ISO 532-3, 2019), although the model used in the standard includes additional stages to account for the way that loudness is combined across ears (Moore and Glasberg, 2007).

Computationally, the most time-consuming stage of the TVL model is the calculation of the excitation pattern, which is estimated from the short-term spectrum of the sound and is used to calculate instantaneous loudness at 1-ms intervals. The excitation pattern is defined as the output of the auditory filters as a function of centre frequency (Moore and Glasberg, 1983). It is estimated by calculating the outputs of an array of level-dependent auditory filters in response to each component of the input signal (after outer- and middle-ear filtering) (Glasberg and Moore, 1990; Moore *et al.*, 1997). The time taken to calculate instantaneous loudness makes it difficult to implement the TVL model in real time. This paper describes the development and training of a deep neural network (DNN) to speed up the computation of instantaneous loudness, allowing real-time implementation. The DNN was trained to predict instantaneous loudness using a large database of speech sounds and artificial sounds (tones alone, bandpass filtered and notched noises, and tones in white or pink noise), with the predictions of the TVL model as a reference (providing the "correct" answer, specifically the loudness level in phons).

STRUCTURE AND TRAINING OF THE DNN

The stimuli used for training had a sampling rate of 16 kHz. Spectra were initially calculated using a 1024-point discrete Fourier Transform, with successive windows being shifted by 560 samples. Then bins were grouped to form 61 bands with centre frequencies up to 8 kHz, with one bin per band for centre frequencies up to 0.2 kHz and 1/9th-octave wide bands for higher centre frequencies. The limit of 8 kHz was chosen due to the sampling rate of the training material. The magnitude of the spectrum was expressed in decibels.

Both accuracy and computation speed were important considerations when choosing the design of the DNN. The DNN was a multilayer perceptron that consisted of an input layer with 61 units (corresponding to the 61 frequency bands), three hidden layers with 150 units each, and a single output unit with linear activation. The output of the DNN was a single loudness level in phon. This was chosen because of its similarity to the input scale. Both scales range roughly from 0 to 110, and the just noticeable difference in loudness is roughly constant on these scales. This facilitated the DNN in developing the mapping from input to output without the need for scale transformations. Simple "rectified linear unit" activations (Nair and Hinton, 2010) were used. Alternative architectures were also considered. Convolutional neural networks did not achieve the same accuracy, probably because the input scale used (logarithmic frequency) did not allow the network to simulate filters that were valid over the whole range of the $\mathrm{ERB_N}$-number scale that is used in the TVL model.

The DNN was optimized with regard to the root-mean-square (RMS) error from the predictions of the TVL model. The Adam optimizer (Kingma and Ba, 2014) was used with its default parameters. All weights were initialized randomly. Three sets of training data were used. First, 500,000 spectra were calculated from the LibriSpeech corpus (Panayotov *et al.*, 2015) from the "clean" development set. The sounds were scaled to have an RMS level of 60 dB SPL. Second, about 700,000 pure tones with levels ranging from 15 to 110 dB SPL and various levels of background noise (from inaudible up to 10 dB below the level of the pure tone) were generated. Third, about 500,000 spectra of bandpass filtered noises and noises with spectral notches were generated. They had various overall levels, bandwidths, notch widths and spectral gradients. To check for "over-fitting", the performance of the DNN was assessed after training for 220, 1000 and 5000 epochs, where an epoch is a complete pass over the entire dataset one time.

ASSESSMENT OF THE DNN

Predictions for speech and everyday sounds

Loudness was predicted for two further sets of data from the LibriSpeech corpus, "clean" set and "other" set (not used for training). Each of them consisted of 500,000 spectra and they were scaled to have an RMS level of 60 dB SPL. Loudness was also predicted for 250,000 spectra derived from the ESC-50 corpus (Piczak, 2015). This corpus contains 50 categories of environmental sounds, for example rain, animals, aircraft, keyboard typing or a washing machine. The sounds were again scaled to have an RMS level of 60 dB SPL. Finally, loudness was predicted for 100,000 spectra from 20 popular songs of the 1960s, which were scaled to have an RMS level of 70 dB SPL. Table 1 shows the RMS error in phons between the predictions of the TVL model and the predictions of the DNN after training for 220, 1000, and 5000 epochs. After 1000 epochs, the RMS error was below 0.5 phons for all classes of sounds. After 5000 epochs the RMS error increased slightly for the LibriSpeech "other" and ESC-50 sounds, which is a sign of "over-fitting". Therefore, in what follows, we focus on the results achieved after training for 1000 epochs.

	Number of epochs		
Test material	220	1000	5000
LibriSpeech "clean"	0.35	0.27	0.28
LibriSpeech "other"	0.55	0.45	0.47
ESC-50	0.56	0.45	0.47
Songs from the 1960s	0.38	0.35	0.31

Table 1: RMS error in phons between the predictions of the TVL model and the predictions of the DNN for sounds not used for training. The error did not vary systematically with the predicted loudness level and the errors had a Gaussian distribution.

Predictions for pure tones

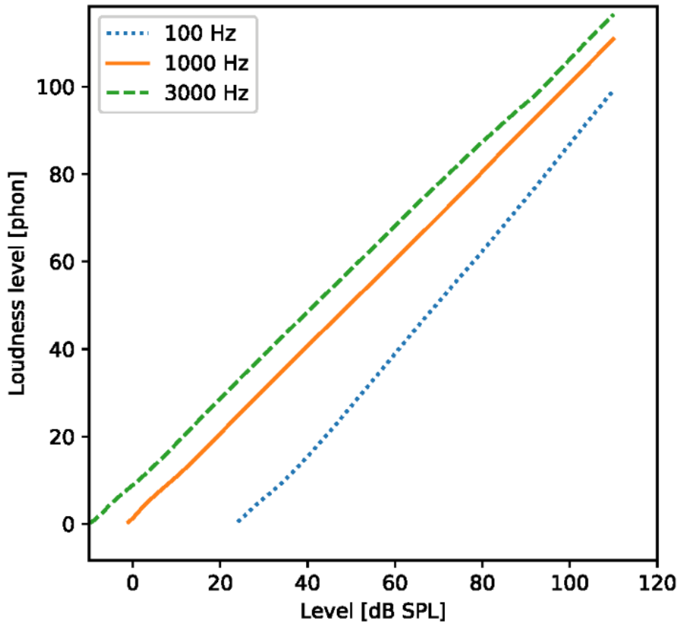

Fig. 2: Loudness level in phons predicted by the DNN as a function of sound level for pure tones with frequencies of 100, 1000, and 3000 Hz.

Figure 2 shows loudness levels predicted by the DNN for pure tones in quiet as a function of input sound level for frequencies of 100 (dotted line), 1000 (solid line) and 3000 (dashed line) Hz, assuming free-field presentation with frontal incidence. The predictions are consistent with empirical data (Hellman, 1976) and are almost identical to the predictions of the TVL model. For the 1000-Hz tone, by definition the loudness level in phons is equal to the physical level in dB SPL. The predictions of the DNN show this relationship almost exactly. The loudness level is greater for the 3000-Hz than for the 1000-Hz tone because 3000 Hz is close to the resonant frequency of the ear canal, so the sound level at the eardrum is boosted relative to that in free field (Shaw and Vaillancourt, 1985). The loudness level is lower at 100 Hz than at 1000 Hz partly because of the attenuation characteristic of the middle ear and partly because less gain is applied by the active mechanism in the cochlea at low frequencies (Cooper, 2004; Moore *et al.*, 1997). Both of these effects are simulated in the TVL model.

Predictions for noises as a function of bandwidth

Figure 3 shows the loudness level of bandpass filtered pink noise centred at 1 kHz, plotted as a function of bandwidth, as predicted by the TVL model and by the DNN. For small bandwidths, the loudness level predicted by the DNN is slightly below that predicted by the TVL model. The predictions of the DNN are actually more consistent with recent empirical data on the loudness of narrowband sounds (Hots *et al.*, 2013), although this is probably just a coincidence.

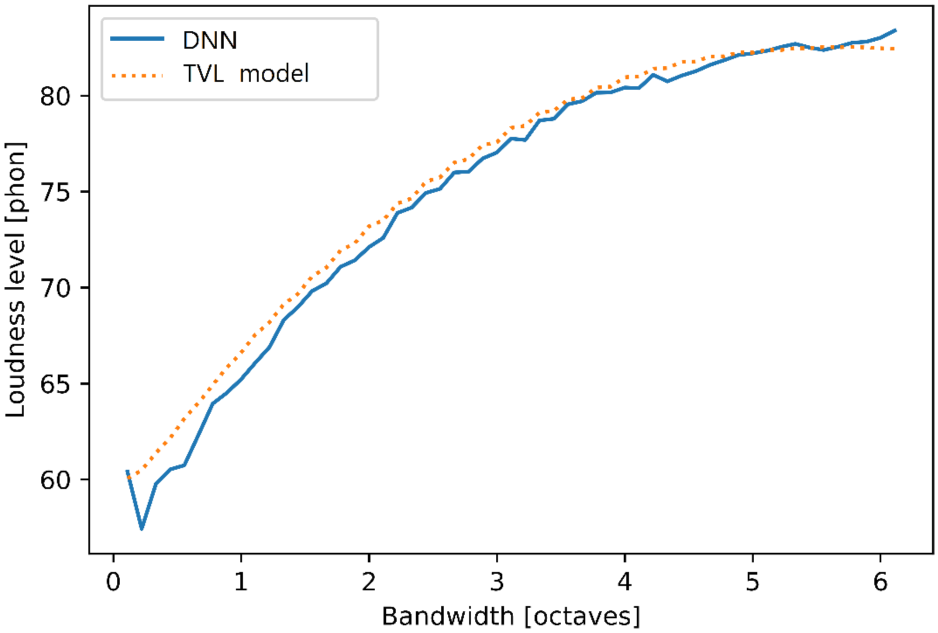

Fig. 3: Loudness level in phons predicted by the DNN (solid line) and by the TVL model (dashed line) as a function of bandwidth.

Note that the sounds whose loudness is predicted in Figures 2 and 3 were included in the sounds used during training, so the accuracy of the predictions is not surprising. Nevertheless, the results show that the inclusion of speech sounds in the training material did not adversely affect the accuracy of the predictions for the artificial tone and noise sounds.

DISCUSSION

The predictions of the DNN for the environmental sounds and music are remarkably accurate. This is noteworthy, since the DNN was trained only using speech and synthetic sounds. This suggests that the DNN generalises well to real-world sounds, and would do so for sounds other than those tested here. The predictions for music were accurate despite the fact that the music test sounds were scaled to have an RMS level of 70 dB SPL, which is higher than the level of 60 dB SPL that was used with the speech sounds used for training. This shows that the DNN works well for sounds with levels that it was not exposed to frequently during training. The good generalisation across level and across types of sounds is probably a consequence of training using tones and noises with a wide range of levels and frequencies as well as with speech sounds. It might be possible to achieve even better generalisation by using an adversarial approach (Szegedy *et al.*, 2013), in which a second DNN tries to find sounds for which the predictions of the first DNN are inaccurate, with the first DNN then adapting in order to achieve more accurate predictions for the problematic sounds. We leave this for a future study.

The gain in computation speed of the DNN relative to the TVL model was a factor greater than 100. This would allow real-time implementation. It is possible with a modern PC (with Intel i7 6th generation central processing unit) to analyse a 24-hour recording at 1-ms intervals in a few minutes.

Potential applications of the DNN include development of a real-time loudness meter and real-time control of levels in broadcasting to ensure (among other things) that the advertisements are not louder than the main programme material (Moore *et al.*, 2003). The DNN could be extended to predict loudness for people with hearing loss (Moore and Glasberg, 1997; 2004). In principle this could be used for on-line control of loudness in hearing aids so as to restore loudness perception more nearly to normal (Launer and Moore, 2003)

CONCLUSIONS

The DNN gave accurate predictions of loudness for environmental sounds and music despite training using speech and synthetic sounds only. This shows good generalisation and suggests that the DNN will give reasonably accurate predictions for a wide variety of everyday sounds. Most predictions were accurate within 0.5 phons, a difference in loudness level that would not be detectable. The DNN calculates instantaneous loudness more than 100 times faster than the TVL model, making real-time implementation possible. This opens up potential applications in broadcasting and in the on-line control of loudness in hearing aids.

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council (UK, grant number RG78536). We thank Brian Glasberg for his collaboration in the development of the TVL model.

REFERENCES

Aibara, R., Welsh, J. T., Puria, S., and Goode, R. L. (**2001**). "Human middle-ear sound transfer function and cochlear input impedance," Hear. Res. **152**, 100-109. doi: 10.1016/S0378-5955(00)00240-9

Cooper, N. P. (**2004**). "Compression in the peripheral auditory system," in *Compression: From Cochlea to Cochlear Implants*, edited by S. P. Bacon, R. R. Fay, and A. N. Popper (Springer, New York), 18-61.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103-138. doi: doi.org/10.1016/0378-5955(90)90170-T

Glasberg, B. R., and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds," J. Audio Eng. Soc. **50**, 331-342.

Hellman, R. P. (**1976**). "Growth of loudness at 1000 and 3000 Hz," J. Acoust. Soc. Am. **60**, 672-679. doi: 10.1121/1.381138

Hots, J., Rennies, J., and Verhey, J. L. (**2013**). "Loudness of sounds with a subcritical bandwidth: A challenge to current loudness models?," J. Acoust. Soc. Am. **134**, EL334-339. doi: 10.1121/1.4820466

ISO 532-3 (**2019**). *Acoustics - Methods for calculating loudness - Part 3: Moore-Glasberg-Schlittenlacher method for time varying sounds* (International Organization for Standardization, Geneva), (draft).

Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980

Launer, S., and Moore, B. C. J. (**2003**). "Use of a loudness model for hearing aid fitting. V. On-line gain control in a digital hearing aid," Int. J. Audiol. **42**, 262-273. doi: 10.3109/14992020309078345

Marshall, A., and Davies, P. (**2007**). "A semantic differential study of low amplitude supersonic aircraft noise and other transient sounds," in *International Congress on Acoustics* (Madrid), pp. 1-6.

Moore, B. C. J. (**2012**). *An Introduction to the Psychology of Hearing, 6th Ed.* (Brill, Leiden, The Netherlands), 1-441.

Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750-753. doi: 10.1121/1.4955005

Moore, B. C. J., and Glasberg, B. R. (**1997**). "A model of loudness perception applied to cochlear hearing loss," Auditory Neurosci. **3**, 289-311.

Moore, B. C. J., and Glasberg, B. R. (**2004**). “A revised model of loudness perception applied to cochlear hearing loss,” Hear. Res. **188**, 70-88. doi: 10.1016/S0378-5955(03)00347-2

Moore, B. C. J., and Glasberg, B. R. (**2007**). “Modeling binaural loudness,” J. Acoust. Soc. Am. **121**, 1604-1612. doi: 10.1121/1.2431331

Moore, B. C. J., and Oxenham, A. J. (**1998**). “Psychoacoustic consequences of compression in the peripheral auditory system,” Psych. Rev. **105**, 108-124. doi: 10.1037/0033-295X.105.1.108

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**, 224-240.

Moore, B. C. J., Glasberg, B. R., and Stone, M. A. (**2003**). “Why are commercials so loud? - Perception and modeling of the loudness of amplitude-compressed speech,” J. Audio Eng. Soc. **51**, 1123-1132.

Nair, V., and Hinton, G. E. (**2010**). “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, edited by J. Fürnkranz, and T. Joachims (Haifa, Israel), pp. 807-814.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). “Librispeech: an ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Brisbane, Australia), pp. 5206-5210.

Piczak, K. J. (**2015**). “ESC: Dataset for environmental sound classification,” in *Proceedings of the 23rd ACM International Conference on Multimedia* (ACM, Brisbane, Australia), pp. 1015-1018.

Rennies, J., Wächtler, M., Hots, J., and Verhey, J. (**2015**). “Spectro-temporal characteristics affecting the loudness of technical sounds: data and model predictions,” Acta Acust. united Ac. **101**, 1145-1156. doi: 10.3813/AAA.918907

Shaw, E. A., and Vaillancourt, M. M. (**1985**). “Transformation of sound-pressure level from the free field to the eardrum presented in numerical form,” J. Acoust. Soc. Am. **78**, 1120-1123. doi: 10.1121/1.393035

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (**2013**). “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199

Thwaites, A., Glasberg, B. R., Nimmo-Smith, I., Marslen-Wilsen, W. D., and Moore, B. C. J. (**2016**). “Representation of instantaneous and short-term loudness in the human cortex,” Front. Neurosci. **10,** article 183, 1-11. doi: 10.3389/fnins.2016.00183

Zorila, T.-C., Stylianou, Y., Flanagan, S., and Moore, B. C. J. (**2016**). “Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing,” J. Acoust. Soc. Am. **140**, 402-408. doi: 10.1121/1.4955005

Zwicker, E., and Scharf, B. (**1965**). “A model of loudness summation,” Psych. Rev. **72**, 3-26. doi: 10.1037/h0021703