DETECTION OF CLICKS IN ANALOG RECORDS USING PERIPHERAL-EAR MODEL

František Rund

Department of Radioelectronics,
Czech Technical University in Prague
Prague, Czech Republic
xrund@fel.cvut.cz

Václav Vencovský

Department of Radioelectronics,
Czech Technical University in Prague
Prague, Czech Republic
vaclav.vencovsky@gmail.com

Jaroslav Bouše

Department of Radioelectronics,
Czech Technical University in Prague
Prague, Czech Republic
bousejar@fel.cvut.cz

ABSTRACT

This study describes a system which detects clicks in sound (audible degradations). The system is based on a computational model of the peripheral ear. In order to train and verify the system, a listening test was conducted using 89 short samples of analog (vinyl) records. The samples contained singing voice, music (rock'n'roll), or both. We randomly chose 30 samples from the set and used it to train the system; then we tested the system using the 59 remaining samples. The system performance expressed as a percentage of correct detections (78.1%) and false alarms (3.9%) is promising.

1. INTRODUCTION

Any undesirable changes in the audio signal are considered as its degradation. According to [1] the degradations can be classified into two groups: global degradations (e.g. background noise, nonlinear distortion, wow and flutter), and localized degradations. Localized degradations are discontinuities in the waveform present only in some samples, such as impulse noise (clicks, crackles, pops, ticks etc.) In this study, we used the term "click" to classify the localized degradations perceived by the listeners as the characteristic noise which is mainly associated with vinyl records. These degradations very often occur in analog records, for example, in historical records, or as a result of damage during the manufacturing process.

Detection of clicks is a principal part of the restoration process (e.g. [1]) or can be used for the audio quality assessment, for example, output quality check during the manufacturing process of the audio records. Manufactures, such as GZmedia [2], usually perform quality control by listening tests with trained employees, however, it is very cost and time demanding. The existing algorithms for impulse detection are based on time domain modeling of the signal (e.g. [1, 3, 4, 5]), or on wavelet transform approach (e.g. [6, 7]). In all of the aforementioned approaches, a detection threshold has to be set. The choice of the threshold is very often empirical, depends on the type of the signal, parameters of the algorithm, etc. Inappropriate threshold setting leads to false detection or missed clicks.

As stated in [1], it is necessary to focus mainly on perceptible degradations. Therefore in this study, we propose a click-detection system based on a computational model of the peripheral ear. The peripheral ear model consists of a physical cochlear model which we previously used for other purpose related with perception [8]. We trained and tested the system by using 89 short sound samples of real music, which contained perceptible clicks. The music samples, which were provided by the vinyl record manufacturer GZmedia, were extracted from four different songs of rock'n'roll music. We conducted a listening test in order to measure the presence of audible clicks.

2. SYSTEM BASED ON A PHYSICAL COCHLEAR MODEL

Figure 1 shows a diagram of the system. The first two blocks represent algorithms simulating the function of the peripheral ear. The remaining blocks process the output signal of the peripheral ear model – this signal is called "internal representation" of the analyzed sound – and give the answer whether the sound contains audible click(s) and temporal position of the click(s) in the signal.

The model of the peripheral ear is composed of two parts, both adapted from the literature. The first part simulates the transformation of the acoustic wave at the entrance of the outer ear into the vibrations of the stapes (the input of the inner ear). The model was adapted from the system called Matlab Auditory Periphery [9]. Resonances of the outer-ear canal are modeled by two parallel 1st-order Butterworth bandpass filters: the first with a gain of 10 dB, lower cutoff frequency of 2.5 kHz, and higher cutoff frequency of 4 kHz; and the second with a gain of 25 dB, lower cutoff

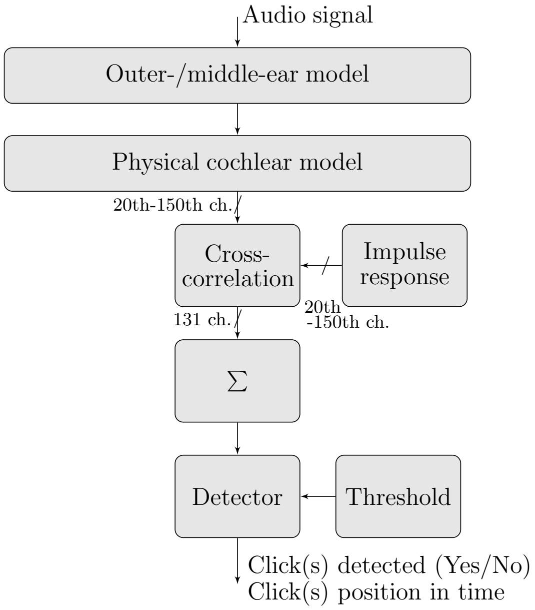

Figure 1: *Diagram of the click detection system*

frequency of 2.5 kHz, and higher cutoff frequency of 7 kHz. The input acoustical signal is first filtered by the two parallel band-pass filters. Then the both filtered signals and the input signal are summed together. The created signal is then processed by a middle ear model transforming it into the stapes displacement. The middle-ear model is composed of two cascaded first order Butterworth filters: a high-pass filter with a cutoff frequency of 50 Hz, and a low-pass filter with a cutoff frequency of 1 kHz. The signal at the output of the second filter is then multiplied by a constant of 45×10^{-9}, which transforms the signal into the displacement (in meters) of the stapes.

The second part of the peripheral ear model transforms the stapes vibrations into the vibrations of the longitudinal segments of the basilar membrane inside the cochlea. The cochlea conducts spatial-frequency analysis – the high frequencies of the incoming sound excite the basilar membrane near the basal site, whereas the low frequencies near the opposite (apical) site. In this paper, the function of the cochlea is simulated by a physical cochlear model described in [10]. The model approximates the basilar membrane by an array of 300 oscillators coupled via surrounding fluid. Therefore the model output is a multichannel signal – the signal in each channel represents displacement of one oscillator. This displacement ξ_i of the i-th oscillator is given by

$$\begin{aligned} m_i\ddot{\xi}_i(t) + h_i\dot{\xi}_i(t) + s_i[2\dot{\xi}_i(t) - \dot{\xi}_{i-1}(t) \\ - \dot{\xi}_{i+1}(t)] + k_i\xi_i(t) = f_{H_i}(t) + f_{\mathrm{OHC}_i}[\eta_i(t)], \end{aligned} \quad (1)$$

where m_i, h_i, s_i and k_i are mass, positional viscosity, sharing viscosity and stiffness of the basilar membrane, respectively. Each oscillator is driven by force $f_{H_i}(t)$ given by

$$f_{H_i}(t) = -G_{S_i} a_{S_i}(t) - \sum_{j=1}^{N} G_i^j \ddot{\xi}_j(t), \quad (2)$$

where $a_{S_i}(t)$ is acceleration of the stapes, $\ddot{\xi}_i(t)$ is acceleration of the oscillators, G_{S_i} and G_i^j are transfer functions obtained by solving wave equations. The second force term f_{OHC_i} represents the cochlear amplifier. In the model, the tectorial membrane connected with stereocilia of the outer hair cells is simulated by another array of oscillators. The stereocilia deflection η_i is given by the differential equation

$$\bar{m}_i\ddot{\eta}_i(t) + \bar{h}_i\dot{\eta}_i(t) + \bar{k}_i\eta(t) = -D_i\ddot{\xi}_i(t), \quad (3)$$

where $\bar{m}_i$ is mass, $\bar{h}_i$ is viscous damping and $\bar{k}_i$ is stiffness, and D_i is a constant. The active force f_{OHC_i} is then calculated from η_i, which is first transformed by a sigmoidal nonlinear function, which attenuates the amplification at high intensities [10]. The model parameters are same as those used in [8]. The characteristic frequencies (CFs) of the model channels, i.e., the frequencies of 10-dB pure tone causing the highest excitation in the given channel, were distributed roughly between 30 Hz and 17 kHz. As well as in [10, 8], the model was implemented in the time domain using the implicit Euler method. The accuracy of this method depends on the sampling frequency – rises with increasing sampling frequency. Therefore by assuming that input stimuli have a sampling frequency at least 44.1 kHz, the input signal to the cochlear model was 10-times upsampled before the processing and then the output signal in each model channel was 10-times downsampled.

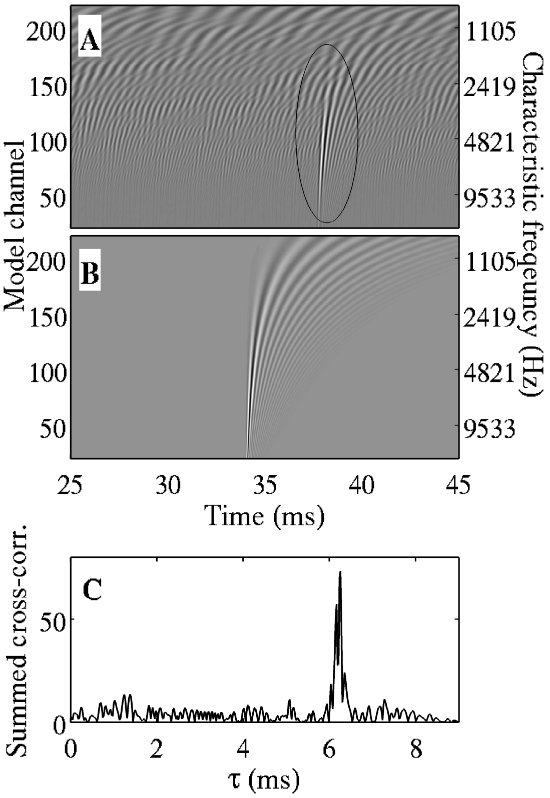

Figure 2: *(A) The auditory model response (internal representation) to a musical signal degraded by a click. The internal representation of the click is marked by the ellipse. (B) The internal representation of an impulse with an amplitude of 2 Pa. (C) Absolute value of summed cross-correlation between the internal representations shown in panels A and B. The cross-correlations were summed across the 20th to 150th model channel (as is indicated in Figure 1).*

Figure 2, panel A shows the internal representation (output signal of the peripheral ear model) of a music sound sample which contains a click. The click created a very distinct pattern in the internal representation, which is easily visible especially at high CFs (above about 2.5 kHz); the pattern is in panel A marked by the ellipse. This internal representation is very similar to the internal representation of an impulse (generated as unit impulse with an amplitude of 2 Pa), which is shown in Figure 2, panel B. Therefore the system detects click(s) in sound samples by calculating cross-correlation of the internal representations obtained in response to an analyzed sound sample and to an impulse. The cross-correlation is calculated between the internal representations in corresponding channels. Then the cross-correlations in the individual channels are summed, which creates a signal with a distinct peak indicating a presence of click. Figure 2, panel C shows the absolute value of the summed cross-correlation between the internal representations shown in the same figure. The cross-correlations were summed between the 20th and 150th channels; the CFs of these channels were 14189 Hz and 2419 Hz, respectively. We have chosen this range experimentally. Notice that the channel numbering is inverted – the first channel has the highest CF (see Fig. 2). This order is given by the physical cochlear model [10] in which the first channel simulates the segment of the basilar membrane which is closest to the stapes (input to the cochlea); this segment has the highest CF. The designed system then detects the presence of click(s) by comparing the amplitude of the summed cross-correlation with a previously defined threshold value; the threshold value used in the system was set during the system training described in the next section. If the amplitude is higher than the threshold value, the system detects click(s) (see Figure 1).

Using this method also allows accurate detection of click position in the time domain. This information could be used for a click removal. A disadvantage of the proposed method is that it is computationally demanding, especially the cochlear model. Therefore it cannot be used in real time.

3. LISTENING TEST, SYSTEM TRAINING AND TESTING

In order to measure the presence of click(s) in sound samples, we conducted a listening test. Since the system compares the summed cross-correlation with a previously defined threshold value, it was necessary to set the threshold value by using some of the results of the listening test (to train the system). The remaining results were then used to test the system.

3.1. Listening test

3.1.1. Stimuli

The stimuli were 800 ms long musical samples shaped on its onset and offset with 80 ms long raised cosine ramps. The samples were extracted from wav files with four different songs; the wav files were converted from analog (vinyl) records containing impulse degradations. The samples contained a singing voice, music (rock'n'roll), or both. All the samples were provided by the vinyl record manufacturer GZmedia. The level of the individual samples was not scaled to the same value for all the samples in order to preserve the dynamic range of the music; the samples were presented with a sound pressure level given by its content – if it contained a 1-kHz pure tone with maximum possible amplitude in the wav file, the presented sound pressure level was 94 dB. For further description about the calibration procedure please refer to [11].

3.1.2. Listeners

Four (all males) listeners aged between 24–33 years participated in the experiment. All of them had normal hearing according to their pure tone hearing thresholds – the thresholds between 250 Hz and 8 kHz were within a range of 20 dB of hearing level. The listeners had no prior experience with this type of experiment.

3.1.3. Procedure and equipment

The experiment was divided into three sessions, each session followed by a compulsory pause. The first session provided the listeners a chance to learn the procedure of the experiment. It consisted of 25 stimuli presented three times in a random order. After the first session we asked the listeners whether they had problems with the procedure or with the detection of clicks in the stimuli; none of them rated the experiment or the detection task difficult. The results of the first session were discarded from the evaluation. Finally the second and third session consisted of 89 stimuli presented in random order and the data from these sessions were taken as the results of the listening test.

The listening test was conducted in a sound insulated booth placed in our laboratory. The listeners were sitting in front of a computer monitor (EIZO S2000) and external soundcard (RME Fireface UC), both connected to a computer placed outside of the booth. The listeners controlled the listening test using a mouse and graphical user interface (GUI) programmed in Matlab, and the stimuli were presented via Sennheiser HD 650 headphones. The GUI consisted of three large buttons labeled as "YES", "REPEAT" and "NO". The listeners were presented with a sound sample and asked to press "YES" or "NO" button based on whether they had perceived click(s) in the stimulus. The listeners had a possibility to repeat the presented stimulus as many time as they desired by pressing the "REPEAT" button. After the response and a subsequent 2-second pause, a next stimulus was presented. No feedback was given to the listeners, except the number of the presented sample within the set.

3.2. System training and testing

The overall set of 89 sound samples was randomly divided into a training set with 30 samples and a testing set with 59 samples. During the training, the system was presented with the samples from the training set and the listening tests results of these samples were used to set a threshold value. This threshold value was then used to predict the occurrence of clicks in the samples within the testing set.

As is shown below in Section 4, some of the samples were not rated clearly, for example, fifty percent of the ratings suggested the presence of click(s) and fifty percent not. For the system training it was necessary to set a rule according to which the sample will be claimed to contain click(s). We claimed to contain click(s) those samples which were rated at least in: (1) 75%, or (2) 50% of the all answers across the sessions and listeners. For these samples, we calculated the summed cross-correlations – the cross-correlations were summed between the 20th and 150th model channel; the CFs of these channels were 14189 Hz and 2419 Hz, respectively. We then calculated the maximum absolute value of the summed cross-correlations for each of the sample and then the minimum value of these maxima across the samples. This gave us a threshold value which was then used during the system testing.

After the threshold value was set by using the training sequence, the system performance was evaluated by using the 59 test samples. Based on the results, each sample was labeled either to contain a click or not. The results are given below together with the results of the listening test.

4. RESULTS AND DISCUSSION

Figure 3 shows the results of the listening test and the system predictions. The bars show the percentage of the positive answers – click was detected – across the all listeners and sessions. The circles above each bar indicate those samples for which the click was predicted by the system. The threshold value used in the system was set by using the listening test results on the 30 training sound samples. We assumed that the samples from the training sequence contained a click if the percentage of positive – click was detected – answers was higher or equal to (1) 75%, or to (2) 50%. In fact, both conditions gave the same threshold value (41.5). Therefore the predictions for both conditions are the same. The predictions are shown in Figure 3 by circles.

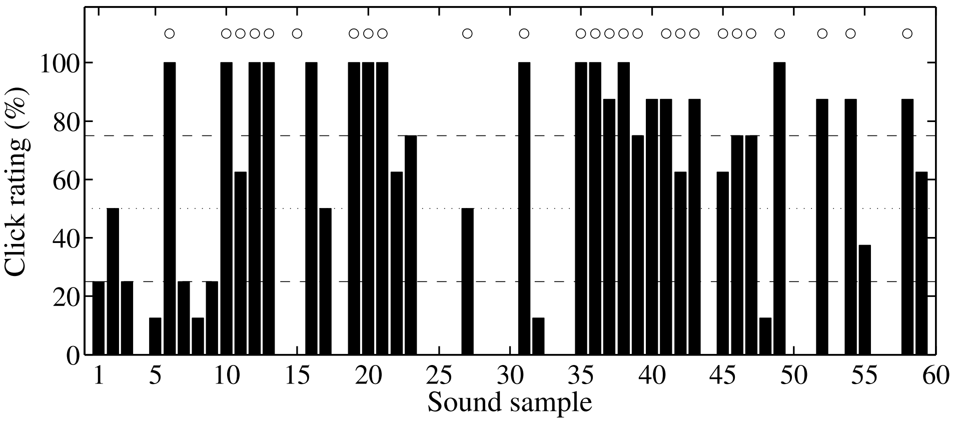

Figure 3: *Results of the listening test and predictions of the click detection system. The bars show the percentage of the click occurrence across the all listeners and responses. Each bar represents one stimulus. Circles indicate the samples in which the system detected a click(s).*

In order to analyze the model performance, we used the "sensitivity measures" for "Yes-No" experiments as is described in [12]. The results of the listening tests were interpreted using two conditions: (1) the samples were marked to contain a click ("Detected") if the rating was equal or higher than 75%, and without a click ("Not detected") if the rating was equal or less than 25% (see the horizontal dashed lines in Figure 3); and (2) the samples were marked "Detected" if the rating was equal or higher than 50%, and "Not detected" if the rating was less than 50% (see the horizontal dotted line in Figure 3). Table 1 shows the calculated model performance. The correct response ("Hit") percentage was calculated by dividing the number of correctly detected clicks by the overall number of samples marked as "Detected" in the listening tests. The "False alarm" percentage was calculated by dividing the number of samples in which the click was predicted and which were marked as "Not detected" by the listeners. The performance of the click detection system is promising; mainly the false alarm rate is very small which may indicate that the threshold value could be smaller to increase the percentage of correct detections.

Table 1: *The system performance.*

	Hit	False alarm
Det. $\geq$75%; Not det. $\leq$ 25%	87.5%	3.9%
Det $\geq$50%; Not det. $<$ 50%	78.1%	3.8%

The clicks were in some of the samples hardly distinguishable from music, e.g., the sound of the plectrum in some samples was very similar to clicks. This contributed to the decreased performance of the detection system. In addition to results of listening tests, in future work we plan to evaluate the system performance against the results of an objective system which uses a reference.

The visual analysis of the model outputs in response to the analyzed sound samples revealed that in some cases the high frequency portion of the model responses may be masked by the musical content. Therefore summing the cross-correlations across a large number of model channels may not be the ideal method for click detection. We plan to focus on these details in future work.

5. CONCLUSIONS

A system allowing to detect the degradation of audio records by clicks (localized degradation caused, for example, by scratches on vinyl records) was designed in this study. The system employs a preprocessing part composed of a computational model of the peripheral ear. The peripheral ear model accounts for the transfer function of the outer and middle-ear, and for the function of the cochlea. The system does not use any reference; it detects clicks by calculating the cross-correlation between the peripheral ear model responses to the analyzed sound and to an impulse. The system sums the cross-correlations across several model channels (at frequencies above about 2.5 kHz) and compares the absolute maximum value with a previously defined threshold value. In order to define the threshold value and then to test the system performance, we used 89 short sound samples (containing signing voice and rock'n'roll music). We evaluated the samples by a listening test in which the listeners were asked whether the samples were degraded by clicks (contained clicks). We randomly chose 30 of the samples for the system training – for setting the threshold value; and then tested the system performance using the remaining 59 samples. The system performance was promising: the correct response rate which depends on the chosen method for the interpretation of the listening test results was higher than 78.1% and the false alarm rate also dependent on the method was smaller than 3.9%. We plan to improve the system accuracy in future work.

6. ACKNOWLEDGMENTS

This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS14/204/OHK3/3T/13. The samples of real damaged records were provided by GZmedia company. Authors thank the volunteers who participated in the listening test, and the anonymous reviewers for comments on the previous version of the manuscript.

7. REFERENCES

[1] S. J. Godsill and P. J. W. Rayner, *Digital audio restoration*, Springer-Verlag, London, 1998.

[2] GZmedia, "Pressing GZ Vinyl," Available at http://www.gzvinyl.com/Manufacturing/Pressing.aspx ?sec=Quality-control, accessed March 10, 2016.

[3] S. J. Godsill and P. J. W. Rayner, "A Bayesian approach to the restoration of degraded audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 267–278, 1995.

[4] Ch. Chandra, M. S. Moore, and S. Mitra, "An efficient method for the removal of impulse noise from speech and audio signals," in *Proc. of the 1998 IEEE Intl. Symposium on Circuits and Systems (ISCAS'98).* IEEE, 1998, vol. 4, pp. 206–208.

[5] M. Niedźwiecki and M. Ciołek, "Sparse vector autoregressive modeling of audio signals and its application to the elimination of impulsive disturbances," *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 1208–1213, 2015.

[6] P. Rajmic and J. Klimek, "Removing crackle from an LP record via wavelet analysis," in *Proc. of the 7th Intl. conf. on digital audio effects (DAFx04)*, 2004, pp. 100–103.

[7] R. C. Nongpiur and D. J. Shpak, "Impulse-noise suppression in speech using the stationary wavelet transform," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 866–879, 2013.

[8] V. Vencovský and F. Rund, "Using a physical cochlear model to predict masker phase effects in hearing-impaired listeners: A role of peripheral compression," *Acta Acustica united with Acustica*, vol. 102, no. 2, pp. 373–382, 2016.

[9] Essex Hearing Research Laboratory, "Auditory modelling at Essex University," Available at http://www.essex.ac.uk/ psychology/department/hearinglab/modelling.html, accessed March 10, 2016.

[10] R. Nobili, A. Vetešnik, L. Turicchia, and F. Mammano, "Otoacoustic emissions from residual oscillations of the cochlear basilar membrane in a human ear model," *Journal of the Association for Research in Otolaryngology*, vol. 4, no. 4, pp. 478–494, 2003.

[11] J. Bouše, "Headphones measurement tool implemented in Matlab," in *Proc. of the 19th International Scientific Student Conf. POSTER 2015*, Prague, 2015, pp. 1–5, Czech Technical University in Prague.

[12] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*, Lawrence Erlbaum associates, publishers, Mahwah, New Jersey, London, 2005.