MARCH 17 2017

Automatic loudness control in short-form content for broadcasting

Leandro da S. Pires; Maurílio N. Vieira; Hani C. Yehia

J. Acoust. Soc. Am. 141, EL287–EL292 (2017)
https://doi.org/10.1121/1.4978023

Automatic loudness control in short-form content for broadcasting

Leandro da S. Pires,[a] Maurílio N. Vieira, and Hani C. Yehia[b]
CEFALA—Center for Research on Speech, Acoustics, Language and Music, Department of Electronic Engineering, Universidade Federal de Minas Gerais, Avenida Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
leandropires@ufmg.br, maurilionunesv@cpdee.ufmg.br, hani@cpdee.ufmg.br

Abstract: During the early years of the International Telecommunication Union (ITU) loudness calculation standard for sound broadcasting [ITU-R (2006), Rec. BS Series, 1770], the need for additional loudness descriptors to evaluate short-form content, such as commercials and live inserts, was identified. This work proposes a loudness control scheme to prevent loudness jumps, which can bother audiences. It employs short-form content audio detection and dynamic range processing methods for the maximum loudness level criteria. Detection is achieved by combining principal component analysis for dimensionality reduction and support vector machines for binary classification. Subsequent processing is based on short-term loudness integrators and Hilbert transformers. The performance was assessed using quality classification metrics and demonstrated through a loudness control example.

[DHC]
Date Received: November 9, 2016 **Date Accepted:** February 19, 2017

1. Introduction

In recent years, loudness control has become one of the most important topics in audio broadcast. The goal of dynamic range processing is to obtain an adequate digital signal level. However, it is commonly misused to obtain songs and advertisements at high volumes (Vickers, 2010). There has been considerable effort toward developing models for loudness assessment of time-varying sounds in broadcast audio (Skovenborg and Nielsen, 2004). In most experiments, classical Zwicker-based spectral loudness summation models (Zwicker, 1977; Moore and Glasberg, 2002) were outperformed by time-averaged root-mean square (RMS) measurements of frequency-weighted signals (Soulodre, 2004). These observations yielded the long-term loudness descriptors presented in the International Telecommunication Union (ITU) recommendation for the broadcasting sector (ITU-R, 2006), based on a weighted power sum of channels.

However, experts have verified that long-term descriptors are insufficient for evaluating loudness jumps in programs owing to short audio content such as live inserts and commercial breaks (EBU, 2010). "Momentary loudness" and "short-term loudness," which use 400 ms and 3 s integrators, respectively, were recommended as additional control criteria (EBU, 2014). We propose a loudness controller for broadcasting that considers these integrators as an additional norm of loudness for TV commercials and applies them in short-form segments detected using a support vector machine (SVM) classifier. After a brief description of the ITU model, the proposed short-form content detector and loudness controller are described in Secs. 2 and 3.

1.1 ITU-R loudness model

The recommended model for broadcasting consists of a sum of psychoacoustically weighted energy measurements per channel. First, the monophonic signals pass through a pre-filter that is designed to include the acoustic effects of the head, whose frequency response is an average of the transfer functions from a speaker to the surface of a rigid sphere over the incident angles of acoustic waves in a multichannel listening environment (ITU-R, 2006). The filtered signals are then weighted using a high-pass filter, which is an approximation of the 70-phon equal loudness contour for high

[a]Also at Agência Nacional de Telecomunicações, Gerência Regional em Minas Gerais, Rua Maranhão 166, 30150-330, Belo Horizonte, MG, Brasil.
[b]Author to whom correspondence should be addressed.

frequencies, with a low end that lies between 70-phon and 40-phon contours (Soulodre, 2004). The combination of both filters is referred to as the *K-weighting filter*.

Fixed correction gains are applied to each channel to compensate for using the same pre-filter in all channels. Then, linear summation yields a composite loudness calculation of the form

$$L_K = -0.691 + 10\log_{10}\sum_i G_i \cdot \frac{1}{T}\sum_{m=0}^{T-1} x_{Ki}^2[m], \tag{1}$$

where $x_{Ki}^2[m]$ is the K-weighted audio signal, G_i represents the fixed gains per channel, T is the duration of the broadcasted audio program, and the constant, -0.691, is a calibration gain such that a 1 kHz full dynamic range test tone corresponds to a fixed reference of 100 phon (Skovenborg and Nielsen, 2004). The absolute logarithmic unit is referred to as the *loudness K-weighted full scale* (LKFS). Relative logarithmic changes are measured in *loudness units* (LUs).

2. Short-form content detector

Similar to neural networks and fuzzy systems, SVMs are labeled as nonparametric classifiers because no *a priori* knowledge of data distributions is assumed. Through supervised learning, they acquire decision functions that classify input data into one of the following two classes: *short-form content* or *program*. Figure 1(a) shows the training and tuning procedures for the proposed audio-based short-form content detector, which is described in Secs. 2.1–2.4.

2.1 Data set

Classification training and validation were carried out using a data set with $m = 129\,685$ training samples, $\mathbf{x}_i$ $(i = 1,\ldots,m)$, that correspond to 150 h of recorded channels (Vyas *et al.*, 2014). Samples were arbitrarily divided into two parts, two-thirds for training and one-third for validation. Each sample is comprised of $n = 10$ audio features. These are the mean and variance of the time-domain measurements of short-term energy and zero-crossing rates, along with frequency-domain features such as spectral roll-off, centroid, and flow (Schuller, 2013).

Principal component analysis (PCA) was employed to reduce feature space and increase the speed of the learning algorithm (Duda *et al.*, 2012). First, the covariance matrix, Σ_x, of the audio features data set was decomposed into singular values to reduce the feature space from n to k dimensions. The smallest value of k was selected such that the quadratic projection error divided by the total variance was less than 1%, and principal components retained 99% of the data variance

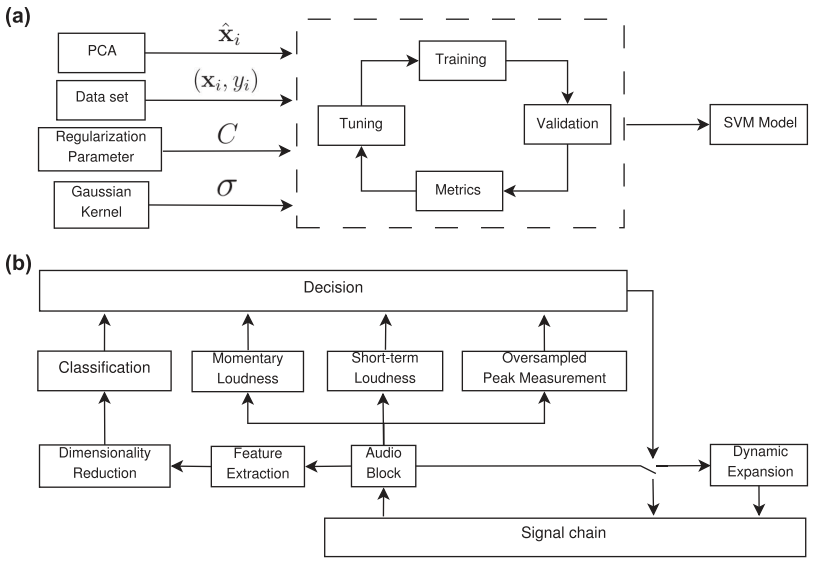

Fig. 1. Block diagrams of the proposed method: (a) process flow for training and tuning the short-form content detector; (b) process flow for audio signal processing and loudness control.

$$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}} \geq 0.99, \tag{2}$$

where S is a diagonal matrix whose non-zero elements are the eigenvalues of Σ_x.

2.2 Tunable parameters

To learn complex decision regions, an SVM classifier maps the original feature space into a higher dimensional space, in which classes can be linearly separated before the minimization of the objective function. This is done using similarity measures computed by radial-basis functions referred to as *kernels* (Abe, 2005). Given m training pairs, $(\mathbf{x}_i, y_i)$, m landmarks, $\mathbf{z}_i$, are set and a new feature vector $\mathbf{f}$ with elements $\mathbf{f}_i = H(\mathbf{x}_i, \mathbf{z}_i)$ is computed, where $H(\cdot,\cdot)$ is the kernel function, and each training sample $\mathbf{x}_i$ is described by a new feature vector with degrees of similarity between $\mathbf{x}_i$ and each landmark $\mathbf{z}_i$. Gaussian kernels were used in the proposed model.

In this work, the following two tunable parameters were identified: the standard deviation of the Gaussian kernel function, σ, and the regularization parameter, C (Abe, 2005). The parameter C penalizes the costs of learning machine parameters so that the model can obtain a regular fit of the data (Duda *et al.*, 2012).

2.3 SVM model

Consider the data set with m training samples, each sample with n audio features and k principal components. Let m k-dimensional training inputs, $\hat{\mathbf{x}}_i\,(i = 1, \ldots, m)$, belong to the class *short-form content*, labeled as $y_i = 1$, or the class *program*, labeled as $y_i = 0$. In addition, let $\mathbf{f}_i$ be the new l-dimensional training inputs obtained after kernel mapping of $\hat{\mathbf{x}}_i$ and $\mathbf{w}$ be the coefficient vector of the separation hyperplane. An objective function of a binary classifier can be formulated as (Abe, 2005)

$$\min_{\mathbf{w}} \left[\sum_{i=1}^{m} y_i \cos t_1\left(\mathbf{w}^T \mathbf{f}_i\right) + (1 - y_i) \cos t_0\left(\mathbf{w}^T \mathbf{f}_i\right) \right] + \frac{1}{2C} \sum_{j=1}^{l} \mathbf{w}_j^2, \tag{3}$$

where for each ith example of the training set, $\mathbf{w}^T\mathbf{f}_i$ is a linear discriminant function in the l-dimensional feature space, and $\cos t_1(\cdot)$ and $\cos t_0(\cdot)$ are the SVM cost functions for the classes $y_i = 1$ and $y_i = 0$, respectively. These cost functions are piecewise linear approximations of log-sigmoid functions. The expression outside the parentheses is the regularization term controlled by the parameter C.

2.4 Training flow

The learning algorithm was trained using different sets of C and σ. For each adjustment, the newly trained classifier was tested on the validation set, and the SVM model with the best scores for a set of performance metrics was selected. The measurement of accuracy was based on true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) in the form

$$\mathrm{Acc}(\%) = 100 \times \left(\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \right). \tag{4}$$

3. Loudness controller

The audio signal processing steps of the proposed loudness controller are shown in Fig. 1(b). The audio segments classified as *short-form content* are dynamically processed if they violate the following maximum level criteria (EBU, 2014): a momentary loudness of −15 LKFS, a short-term loudness of −18 LKFS, and an oversampled peak measurement that is 1 dB below the maximum available peak level or relative to full-scale. Processing experiments were performed on audio tracks that were extracted from captured transport streams broadcasted by terrestrial television stations.

3.1 Loudness integrations

The present work uses the aforementioned descriptors, *short-term loudness* and *momentary loudness*, designed for loudness meter displays. They use a sliding rectangular time window of length T and are computed using Eq. (1), with T corresponding to 3 s and 400 ms (EBU, 2010). The latter has a first-order low-pass smoothing filter at its output for easy reading of sequenced measured values.

Considering a typical broadcast audio sampling frequency of 48 000 samples/s, the classified audio stream was processed in 3 s duration blocks (144 000 samples). Fifteen momentary loudness integrations were computed over sub-blocks of 19 200 samples with 50% overlap. In addition, short-term loudness integration and an oversampled peak measurement were performed over the entire 3 s duration block. Violation of any of the previously mentioned criteria triggered a dynamic expander that controlled the loudness levels. The samples that violated the peak level constraint were subject to minimal attenuation (1 dB) to prevent clipping.

3.2 Dynamic expansion

Traditional approaches for dynamic expanders derive gain curves directly, using the absolute values of samples multiplied by a scalar ratio (Zölzer, 2002). However, modern compressors and expanders employ one-pole smoothing filters in level and peak detection stages to minimize undesirable artifacts due to abrupt change in magnitudes at the gain computing step (Giannoulis *et al.*, 2012). In our proposal, gain curves were derived using a *Hilbert transformer* (Oppenheim and Schafer, 2010). A Hilbert transformer with a real audio signal $x_r[n]$ at its input produces a real output $x_i[n]$, and the analytic signal, $x[n]$, is composed of both signals. Its magnitude, referred to as *instantaneous amplitude*, corresponds to the envelope of the input audio signal $x_r[n]$.

The attenuation applied to the envelope samples whose magnitude is below a threshold level is computed as the ratio of the envelope sample level to the threshold level. The latter is set to be 10 dB below the previously measured peak level, for the following two reasons: (1) 10 dB of level change corresponds approximately to a factor of 2 in loudness perception (Zwicker, 1977), and (2) a crest factor (difference between peak and RMS levels) of 10 dB is the approximate crest factor that would be obtained using an analog tape recording, before current dynamic overcompressed audio (Vickers, 2010).

4. Results and discussion

The effectiveness of the proposed short-form content detector was assessed using the hit percentages on the validation set for the SVM models trained with different values of the regularization parameter, C, and the kernel standard deviation, σ, ranging from 10^{-3} to 10^{3} on a logarithmic scale. The results obtained from four different TV channels are shown in Fig. 2. Note that for narrow kernel standard deviations (small values of σ), strong regularization of the SVM model (small values of C) smoothed the SVM objective function, and resulted in underfitting. However, Fig. 2 shows that as σ increases, accuracy increases with weaker regularization, implying that the model was not prone to overfit the data. In addition, classifiers with or without dimensionality-reduced data achieved the best classification performances for the same (C, σ) pairs, illustrating the importance of using a well-tuned classifier, and implying that the selected features were sufficiently relevant and non-redundant.

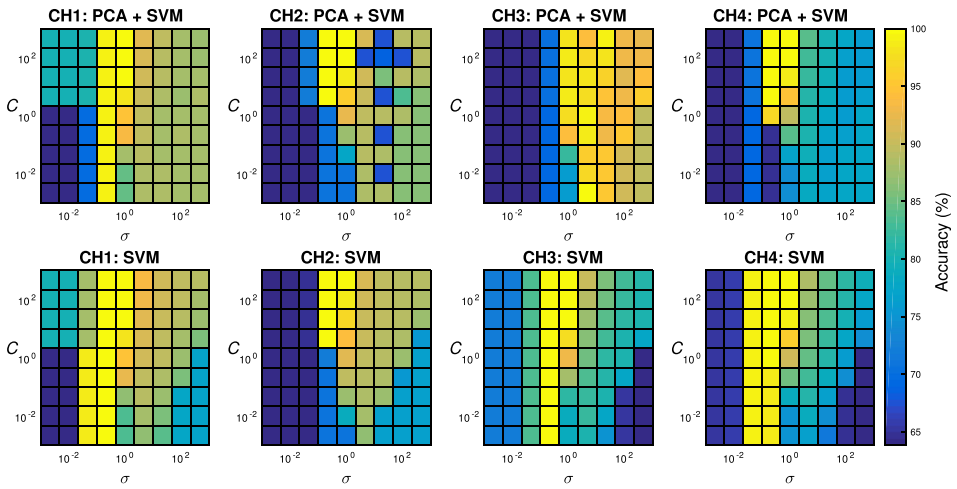

Fig. 2. (Color online) Training and tuning of the short-form detector. Colors indicate hit percentages on validation sets with audio features from four different TV channels. Results for dimensionality-reduced data are shown in the first row. Classifier performances in the original feature space are shown in the second row. Different sets of regularization parameters (C) and kernel standard deviations (σ) yield different classification accuracy values.

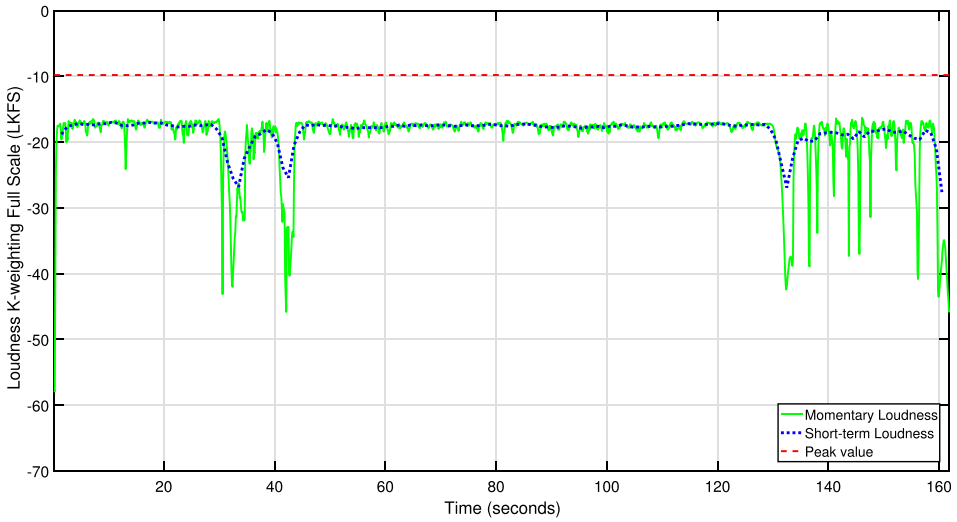

Fig. 3. (Color online) Loudness integrations along an excerpt of a TV program that interchanges dialogue passages with highly compressed songs (Momentary: solid line; Short-term: dotted line). In the latter part, short-term loudness integrations are occasionally greater than −18 LKFS, triggering the loudness controller illustrated in Mm. 1. Dashed line indicates the maximum oversampled peak value.

The significance of the differences in the accuracy between the classifiers with and without PCA was assessed using two one-sided McNemar's statistical tests (Dietterich, 1998), with the null hypothesis rejected at the 95% confidence level when less than seven principal components were used. Therefore, the audio feature space, originally represented by ten explicit audio features, can be represented by its first seven principal components without significant loss of accuracy.

The proposed loudness controller processed audio tracks whose levels violated the adopted criteria, while maintaining the differences between short-term loudness integrations of adjacent blocks at less than 1 LU, for preventing perceivable loudness jumps in the streamed programs. Its effectiveness can be illustrated using an example. Consider the deferred-time loudness integrations of an audio segment shown in Fig. 3. This excerpt is from a TV program that interchanges dialogue passages with highly compressed songs, where it is observed that short-term loudness integrations overshoot the maximum level criterion in some cases. The loudness measurements performed over the input audio and the punctual dynamic expansions of the streamed output waveform can be seen in Mm. 1.

Mm. 1. Loudness control of 160-s audio content. Short-term loudness integrations performed over input/output audio are updated on the top display. Momentary loudness integrations per block are shown in the array plot on the right. Dynamic expansions are displayed on the time scope on the left. Sound is suppressed owing to reproduction rights. This is a file of the type "mp4" (577 kB).

Commercial block detectors that use audiovisual features may build more complex decision regions (Vyas *et al.*, 2014). One of the challenges of this work was to determine whether satisfactory results could be achieved with a limited number of audio features, in terms of developing a classifier that does not introduce significant delay. In future work, the impact of introducing additional low redundancy audio features on classification performance should be evaluated. In addition, learning machine performance can be improved by training the machine using samples from the broadcaster's programs that it is expected to classify.

For the loudness controller, one concern was about choosing the processing block size. The chosen block corresponds to 3 s of audio, which makes short-term loudness integration time convenient, in addition being a comfortable limit for live processing. Thus, the proposed method can be used by small stations that lack dedicated audio processing hardware, or typically run untreated advertisements during their playout schedules.

Acknowledgments

The authors thank UFMG/PRPq, CNPq, CAPES, FAPEMIG, and Anatel for funding this study.

References and links

Abe, S. (**2005**). *Support Vector Machines for Pattern Classification* (Springer, London, UK).

Dietterich, T. G. (**1998**). "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comp. **10**, 1895–1923.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2012**). *Pattern Classification* (John Wiley & Sons, Hoboken, NJ).

EBU (**2010**). Tech 3341, "'EBU Mode' metering to supplement loudness normalisation" (European Broadcast Union, Geneva).

EBU (**2014**). R128-s1-2016, "Loudness parameters for short-form content (advertisements, promos, etc.)" (European Broadcast Union, Geneva).

Giannoulis, D., Massberg, M., and Reiss, J. D. (**2012**). "Digital dynamic range compressor design—A tutorial and analysis," J. Audio Eng. Soc. **60**, 399–408.

ITU-R (**2006**). BS. 1770, "Algorithms to measure audio programme loudness and true-peak audio level" (International Telecommunications Union, Geneva).

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

Oppenheim, A. V., and Schafer, R. W. (**2010**). *Discrete-time Signal Processing*, 3rd. ed. (Pearson Higher Education, New York).

Schuller, B. W. (**2013**). *Intelligent Audio Analysis* (Springer, New York).

Skovenborg, E., and Nielsen, S. H. (**2004**). "Evaluation of different loudness models with music and speech material," in *Audio Engineering Society Convention 117*, San Francisco, CA.

Soulodre, G. A. (**2004**). "Evaluation of objective loudness meters," in *Audio Engineering Society Convention 116*, Berlin, Germany.

Vickers, E. (**2010**). "The loudness war: Background, speculation, and recommendations," in *Audio Engineering Society Convention 129*, San Francisco, CA.

Vyas, A., Kannao, R., Bhargava, V., and Guha, P. (**2014**). "Commercial block detection in broadcast news videos," in *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, ACM, p. 63.

Zölzer, U., ed. (**2002**). *DAFX: Digital Audio Effects*, Vol. 1 (John Wiley & Sons, Hoboken, NJ).

Zwicker, E. (**1977**). "Procedure for calculating loudness of temporally variable sounds," J. Acoust. Soc. Am. **62**, 675–682.