# CONFIDENCE MEASURES FOR ACOUSTIC DETECTION OF FILM SLATES BASED ON TIME-DOMAIN FEATURES

*Markus S. Schlosser*

Deutsche Thomson OHG
Karl-Wiechert-Allee 74,
30625 Hannover, GERMANY
markus.schlosser@thomson.net

**ABSTRACT**

An acoustic detector for film slates is proposed to assist a human operator with the synchronization of audio and video in post-production. To be computationally efficient, the signal analysis is restricted to time-domain features. Although the features are statistically dependent, separate classifiers are trained for each of them. The statistical dependence is taken into account during the combination of the log-likelihood ratios provided by the individual classifiers. The overall confidence in a classification is determined as a weighted sum of the individual log-likelihood ratios, where the weights depend on the correlation between the different features. Experimental results for real-world recordings from film sets show that the confidence measures allow for a fast identification of the film slates while minimizing the interference from false detections.

***Index Terms***— Acoustic signal detection, reliability estimation, time domain analysis, feature extraction

## 1. INTRODUCTION

For film, sound is almost always recorded separately from the images, in a so-called double-system [1, 2]. Therefore, audio and video need to be synchronized during post-production. A sophisticated technical solution for this task consists in using in-camera timecode, which allows for an automatic synchronization. The audio recorder sends its timecode to the camera, where it is optically recorded along with the images. In the future, digital film cameras may even record both video and audio in a single system.

It is, however, still very common to use a simple slate (also called clapperboard or sticks) at the set followed by a manual inspection of the audio and video material by human operators in post-production. For this case, the presented acoustic slate detector aims at speeding up the process of synchronizing audio and video by automating the identification of acoustic events that result from the clapping of a slate.

This task represents a special case of the identification of transient sounds. Similar systems are applied in the classification of impact sounds [3], transient sonar sounds [4], percussive instruments [5], and even transient phenomena on power lines [6]. The short duration of transient sounds makes their classification challenging. In our case, additional difficulties consist in the recording conditions varying widely for different film sets and only little training data being available due to the restrictive policies encountered in film making.

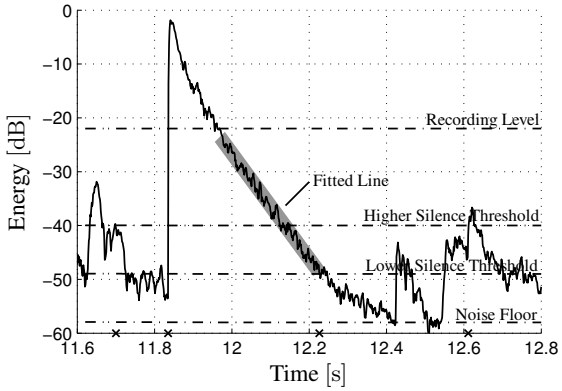


**Fig. 1**. Example of an energy envelope for a true slate (the 'x-marks on the time axis indicate the extracted time events evaluated by the algorithm)

The acoustic slate detector consists of two stages: identification and evaluation of candidates. The identification makes use of the fact that slates are highly percussive sounds (see also Fig. 1). During the subsequent evaluation, each candidate is assigned a confidence measure. This allows operators to start with the most prospective candidates and to discard all remaining false detections upon finding the slate - even if these false detections have been assigned rather high confidence values themselves.

The paper is outlined as follows: Sect. 2 describes the identification of candidates and Sect. 3 their evaluation. Experimental results are presented in Sect. 4, which lead to the conclusions in Sect. 5.

## 2. IDENTIFICATION OF CANDIDATES

As slates are highly percussive sounds, the identification of prospective candidates is rather simple and can be done using one of the many algorithms that exist to detect percussive onsets [7–9]. We simply detect strong energy increases in the high-frequency components of the signal. The specific processing steps are a high-pass filter followed by a calculation of the energy envelope, a derivation, and a peak-picking.

Most noise sources are likely to have a low-pass characteristic whereas transient signals are broadband events. Therefore, the high-pass filter in the first processing step effectively emphasizes transient parts of the signal. It is designed with a broad transition band, so as to gradually increase the weight given to higher frequency components. Furthermore, as the estimated onset of the slate shall be accurate to at least one quarter of a video frame (i.e., $10\,\mathrm{ms}$ at $25$ frames per second), the length of the rectangular window for the calculation of the energy envelope is set to $5\,\mathrm{ms}$ to suppress noise. Finally, the peak-picking uses a global threshold as the sound of the slate closures depends on the overall recording conditions, but it is not correlated with surrounding sounds.

## 3. EVALUATION OF CANDIDATES

The candidates found in the first stage are evaluated based on several time-domain features. A confidence measure is assigned to each candidate using log-likelihood ratios, which are modeled during the training phase.

### 3.1. Feature Extraction

The energy increase detected in the first stage is augmented by several other features extracted from the energy envelope of the original, not high-pass filtered signal. Merely a low-cut filter is applied to suppress very low-frequency noise sources, like e.g., the $50\,\mathrm{Hz}$ or $60\,\mathrm{Hz}$ hum or the mechanical noise of a rolling film camera.

In total, the following ten features are extracted for each candidate (see also Fig. 1): the energy increase, the position and value of the maximum energy, the slope and the Mean Square Error (MSE) of a line fitted to the energy decay, the difference between the measured maximum and one predicted by the fitted line, the occurrence time of the candidate, its duration, and the duration of silence periods before and after the candidate. As a prerequisite for the extraction of several of these features, reliable estimates of the noise floor and the recording level are needed. These estimates are determined based on a histogram of the energy envelope values.

As far as the maximum is concerned, its value is taken relative to the estimated recording level to be robust against changes in the recording conditions. Furthermore, as the position of the maximum is independent of the reverberation time of the room, it is used instead of the temporal centroid of the envelope, which is typically chosen in material classification [3–5]. Finally, the maximum is replaced by an earlier, local maximum if the absolute maximum seems to be due to early reflections.

The line fitting is motivated by the fact that, due to reflections at the walls, the floor, and the ceiling, the energy of the clapping sound trails off exponentially – at least approximately – with the room impulse response. In the logarithmic domain, this results in a linear decrease, which is not only easier to model but also allows for an easier assessment of the quality of fit, e.g., using the proposed MSE.

The exponential decay does not hold for the early, discrete reflections but only for the late, diffuse ones. Therefore, the line fitting is restricted to the final part of the acoustic event. Furthermore, the sampling points are weighted according to their distance to the estimated noise level as the low-energy part is likely to be more corrupted by background noise.

Finally, the energy decrease may be interrupted by simultaneous background noise or even other foreground sounds. In this case, the candidate receives an additional penalty as slates typically appear solo. The strength of the penalty depends on the remaining distance to the silence threshold.

The occurrence time takes care of the fact that slates are typically only clapped at the beginning or end of a recording (normal slates versus tail slates). The duration of a candidate is the time span that the energy envelope stays above the lower silence threshold. Accordingly, the duration of the silence periods before and after the candidate are those time spans that the energy envelope needs to surpass the higher silence threshold again after it has dropped below the lower one. This hysteresis prevents spurious sounds from being detected.

### 3.2. Binary Classification

Based on these $N = 10$ features $x_i$, potential candidates need to be classified into one of the two events: slate ($M_1$) and non-slate ($M_2$). If the probability density functions for the slate and non-slate classes were known, the ratio between the likelihood that a feature vector $\mathbf{x} = \{x_1, \ldots, x_N\}$ is observed for a slate and for a non-slate could simply be used as a confidence measure for this classification [10]. Its logarithm is preferred here due to the nicer value range, i.e.,

$$\mathrm{CM} = \log\left(\frac{f_{\mathbf{X}|M_1}(\mathbf{x}|M_1)P(M_1)}{f_{\mathbf{X}|M_1}(\mathbf{x}|M_2)P(M_2)}\right). \quad (1)$$

A huge amount of training data would be needed to model our ten-dimensional feature space accurately. If the features $x_i$ were statistically independent, Eq. (1) would, however, be equivalent to

$$\mathrm{CM} = \log\left(\frac{P(M_1)}{P(M_2)}\right) + \sum_{i=1}^{N} \log\left(\frac{f_{\mathbf{X}|M_1}(x_i|M_1)}{f_{\mathbf{X}|M_1}(x_i|M_2)}\right). \quad (2)$$

Hence, all features could be modeled separately making the modeling a lot easier. Unfortunately, statistical independence

may only be reasonably assumed for the duration of the silence periods before and after the slate as well as the occurrence time in our case.

A principal component analysis could be applied to render the features at least uncorrelated – which is identical to independent in the Gaussian case [10]. This would, however, still require quite a lot of training data to be able to estimate the cross-correlation matrix accurately. Furthermore, it would be difficult to interpret the transformed features physically and, thus, to include background knowledge during the modeling of the likelihood ratios.

This is especially important to prevent the modeling from being restricted to the recording conditions underlying the training data. For example, the reverberation time, the presence of echoes, and the amount of background noise may vary widely between recordings. Furthermore, especially challenging cases were taken into account as much as possible, like e.g., soft slates clapped in front of an actors face or slates that are recorded from a large distance or through clothing using lavalier microphones. Finally, the probability distributions of some features contain singularities, like e.g., the duration of the silence period before a slate. As the movement of the articulated arm of the slate can make a slight noise itself, its is disproportionately likely that this silence period is non-existent.

### 3.3. Feature Weighting

As a consequence, it was decided to model the log-likelihood ratio for every feature separately despite their statistical dependence. To account for the statistical dependence at least to some extent, weights $w_i$ are assigned to every log-likelihood ratio during the summation

$$\mathrm{CM} \approx \log\left(\frac{P(M_1)}{P(M_2)}\right) + \sum_{i=1}^{N} w_i \log\left(\frac{f_{\mathbf{X}|M_1}(x_i|M_1)}{f_{\mathbf{X}|M_1}(x_i|M_2)}\right) . \tag{3}$$

The weights depend on the correlation between the different features and are determined by the following heuristic rule

$$w_i = \frac{1}{\sum_{j=1}^{N} |\rho_{ij}|^n} , \tag{4}$$

where $\rho_{ij}$ represents the correlation coefficient between the $i^{\text{th}}$ and $j^{\text{th}}$ feature within the slate class and $n$ an adjustable parameter. Outliers surpassing the $3\sigma$-bound are suppressed during the calculation of the correlation coefficients.

This choice makes sure that completely uncorrelated features receive the weight one and that $N$ completely correlated features receive the individual weight $\frac{1}{N}$ so that their total weight again amounts to one. The exponent $n$ indicates how strongly the correlation is taken into account in-between. The bigger the exponent, the smaller the influence of a potential correlation. Therefore, $n$ should be chosen larger if only little data is available as the estimated correlation coefficients are not very accurate. It should be noted that this approach ignores the correlation between the features within the non-slate class. A similar distribution as in the slate class is assumed implicitly.

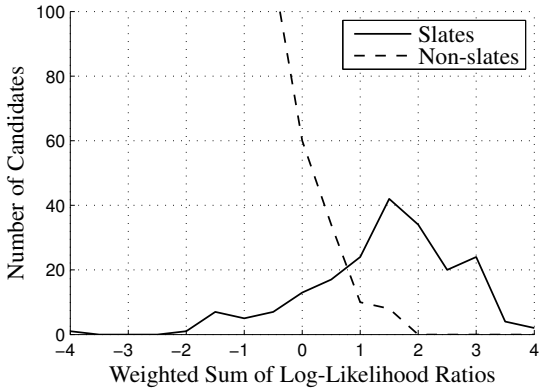


**Fig. 2**. Overall distribution of confidence measures

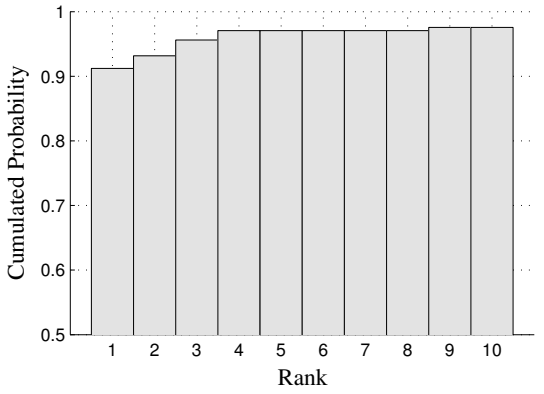


**Fig. 3**. Overall average of ranking achieved by true slates in their respective take

## 4. PERFORMANCE EVALUATION

Material from 4 different film sets with 205 slate closures in total was available to train and evaluate the algorithm. It stems from indoor as well as outdoor recordings. The quality ranged from rather clean to pretty noisy due to, e.g., the mechanical noise of the rolling camera. For one source, many slates are even often obscured by other foreground sounds. Finally, the sampling rate ranged from 11025 Hz to 48 kHz.

The overall results for the complete test material are depicted in Figs. 2 and 3. In Fig. 2, the distribution of the weighted sum of log-likelihood ratios is shown for the slate and non-slate classes. As detailed in Sect. 3.3, these represent the confidence measures. It can be seen that the distributions for the slate and non-slate classes overlap significantly. Using a fixed threshold would either lead to many missed detections

or to a high false alarm rate.

Unfavorable recording conditions cannot only lead to slate sounds being assigned low confidence measures. Other percussive sounds, like e.g., clapping, door slams, or foot steps on high heels, can also sound very similar. Finally, the discrimination becomes even more difficult for exterior recordings as, due to the missing reverberation, the line fitting cannot be used to evaluate the candidates anymore.

Therefore, the detection of the acoustic slate closures cannot be completely automated. The proposed algorithm is, however, able to assist a human operator in locating the slate closures and, thus, in speeding up the synchronization of audio and video significantly. As can be seen in Fig. 3, more than $90\,\%$ of the slates achieve Rank $1$, i.e., they are assigned the highest confidence measure in their respective take. The reason for this is that, if the true candidates are assigned rather low confidence values due to unfavorable recording conditions, the confidence values of the false detections tend to be even lower. In addition to this, there is typically only one slate (per camera) in every take. Consequently, the human operator will immediately be presented the true slate in most cases. Upon finding the slate, all remaining false detections can then be discarded even if they have been assigned quite high confidence measures themselves. This clearly demonstrates the importance of assigning these confidence measures.

Finally, Fig. 3 also shows that, if a slate has not been among the first few candidates, it has likely been assigned a really low confidence measure. This is often due to other foreground sounds being present in parallel, which alters the shape of the energy envelope significantly and which is almost impossible to discern in the time-domain. Additionally, $4$ out of the $205$ slates are so soft or muffled that they are even failed to be detected at all during the identification of prospective candidates in the first stage. As a consequence, if a slate has not been among the first few candidates, the human operator should revert to the conventional way of locating the slate closure by checking the signal envelope or simple "brute force" listening.

On a Pentium IV with $2.8\,\mathrm{GHz}$, the acoustic slate detector runs approximately ten times faster than real-time for stereo signals at a sampling frequency of $48\,\mathrm{kHz}$. This has been achieved by restricting the processing to the time-domain. Furthermore, only the high-pass and low-cut filters and the envelope calculation need to run at the original sampling rate. The formers are specifically designed to be filters of a very low order, and the latter can be implemented efficiently using a moving average filter. As the envelope calculation is equivalent to a low-pass filtering, all subsequent processing steps can be performed on a highly subsampled signal.

## 5. CONCLUSIONS

Although acoustic slate detection is a challenging problem, experimental results show that the proposed algorithm can significantly speed up the synchronization of audio and video in post-production by assisting a human operator in locating the slate closures. For high-quality recordings with clean and unobscured slate closures, these are typically assigned higher confidence measures than any false detection. Under unfavorable recording conditions, the confidence measures tend to be lower, but most slates are, nevertheless, assigned the highest confidence measure in their respective take. This highlights the importance of using these confidence measures. They allow for a detection of almost all slates without too much interference of false detections.

## 6. REFERENCES


[1] T. Holman, *Sound for Film and Television*, Oxford: Focal Press, 2nd edition, 2002.

[2] H. Wyatt and T. Amyes, *Audio Post Production for Television and Film*, Oxford: Focal Press, 3rd edition, 2005.

[3] B. L. Giordano, *Perception in Impact Sounds*, Ph.D. thesis, University of Padova, 2005.

[4] S. Tucker and G. J. Brown, "Classification of transient sonar sounds using perceptually motivated features," *IEEE Journal of Oceanic Engineering*, vol. 30, no. 3, pp. 588–600, 2005.

[5] G. Peeters, S. MacAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7 (2000)," in *Proceedings of International Computer Music Conference (ICMC)*, 2000.

[6] L. Angrisani, P. Daponte, and M. D'Apuzzo, "Wavelet network-based detection and classification of transients," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, no. 5, pp. 1425–35, 2001.

[7] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–47, 2005.

[8] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999, pp. 3089–92.

[9] P. Masri, *Computer Modelling of Sound for Transformation and Synthesis of Musical Signals*, Ph.D. thesis, University of Bristol, 1996.

[10] K. Kroschel, *Nachrichtentheorie: Signal- und Mustererkennung, Parameter- und Signalschätzung*, Berlin: Springer, 2nd edition, 1996.