TOWARDS A SEMANTIC DESCRIPTOR OF SUBJECTIVE INTENSITY IN MUSIC

Vegard Sandvold
NOTAM
Oslo, Norway

Perfecto Herrera
Universitat Pompeu Fabra
Barcelona, Spain

ABSTRACT

In this paper we present a descriptive study of perceived intensity in popular music. We have designed a taxonomy for the purpose, and created an intensity model based on low-level descriptors extracted from an objective data set, which can reliably predict a category label for the "subjective intensity" of most popular music recordings.

1. INTRODUCTION

This paper presents a descriptive study of perceived intensity in popular music, with the goal of establishing an objective model of subjective intensity built from low-level descriptors extracted from the audio data.

Semantic, or high-level descriptors are the focus of several current research projects into MIR, like SIMAC [1] and Semantic HIFI [2] . Automatic extractors of music metadata such as tonality and rhythm are valuable building blocks for advanced music retrieval applications that may provide functionality such as automatic playlist generation and music recommendation. Today, most commercial systems are based on editorial metadata and content descriptions that are manually annotated, either by experts or by a community of users. Content-based retrieval and automatic metadata extraction is a natural next step.

Intensity in music, or *the sensation of energy we get from listening to music*, is a concept commonly used to describe music content. Although intensity has a clear subjective facet, we hypothesize that there exists a basic definition that can be objectively evaluated, i.e. a certain set of salient features determining intensity that most people agree on. In this paper we will establish an objective intensity model by doing the following:

1. Design an intensity taxonomy, where each category is accompanied by textual descriptions that communicate the desired interpretation of subjective intensity to the listener. This provides us with a vocabulary for talking about intensity in music.
2. Show that people perceive intensity in a fairly consistent manner, given the intensity taxonomy. We establish this by means of a listening test, where subjects are asked to assign intensity category labels to excerpts of music recordings.
3. Measure salient features of the audio data using appropriate signal processing that yields low-level descriptors. We will use the terms feature and descriptor interchangeably in this paper.
4. Model intensity by applying a pattern recognition algorithm to the excerpts, represented by feature vectors and the assigned intensity category labels. Unknown music is classified using the resulting intensity model.

A similar study was performed by Zils and Pachet [3], where perceived intensity was divided into four categories named from "low" to "very high". They also performed a listening test, and relied mainly on their Extractor Discovery System (EDS) software to discover new low-level descriptors correlated with intensity. Their model achieved reasonably good accuracy, but provided little insight into how the salient features of intensity interact. Our study reinforces many of their findings, as the data from our listening test is contributed by a different and larger population.

2. METHOD

2.1. Designing an intensity taxonomy

Perceived intensity in music is not a well-defined concept. It is not purely determined by a long-term loudness sensation, but may also be connected to tempo, instrumentation, rhythmic complexity etc. We want, however, to model the sensation of intensity as it is, independent of other concepts. Great care was therefore put into avoiding undesirable connotations in the textual descriptions of the intensity taxonomy, in order not to bias the subjects of our listening test. Words and sentences relating to concepts such as volume, tempo and emotion were specifically avoided.

The taxonomy was defined as follows:

Wild Marked by extreme lack of restraint or control; intensely vivid. Synonyms: *intense, manic, fiery*.

Energetic Possessing or exerting or displaying energy. Synonyms: *lively, sparkling, raucous/rowdy, exciting*.

Moderate Being within reasonable or average limits; not excessive or extreme. Synonyms: *laid-back/mellow*.

[1] http://www.semanticaudio.org/
[2] http://shf.ircam.fr/

Figure 1. Screenshot of RateIt!, our Internet-based listening test.

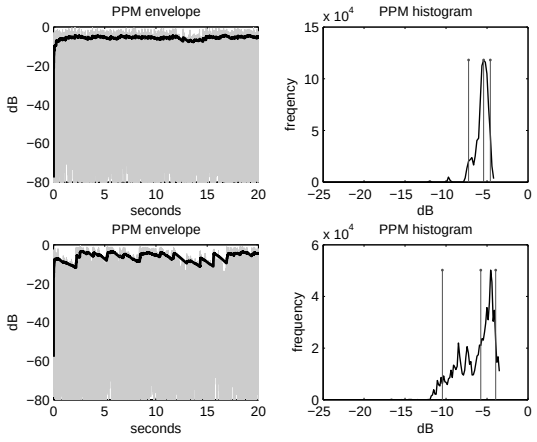

Figure 2. PPM envelopes and histograms for high intensity music (top) and low intensity music (bottom). The median, 5^{th} and 95^{th} percentile are marked onto the histograms.

Soft Having or showing a kindly or tender nature. Synonyms: *gentle, soothing, calm/peaceful*.

Ethereal Characterized by lightness and insubstantiality; as impalpable or intangible as air. Synonyms: *detached, hypnotic, unreal*.

The definitions are taken from WordNet [3], an online lexical reference system. An exception is the definition of *soft*, which has an undesirable connotation related to volume. We used instead the definition of *gentle*, one of the listed synonyms.

The synonyms, and two of the category names, can be found in the mood taxonomy of All Music Guides [4], which also provides lists of reference albums and music titles for each mood. This connection is interesting, as it provides us with a second set of objective data for future evaluations of the final intensity model.

2.2. Gathering objective intensity data

An objective data set of music recording excerpts, sorted into intensity categories, was gathered through a listening test. We implemented a forced-choice style, Internet-based survey requiring only a JavaScript-enabled browser and MP3 playback software of the participants. A screenshot of the survey interface is shown in Figure 1.

In the survey, subjects were presented with a sequence of short excerpts of music and instructed to assign excerpt to one of the five intensity categories. The excerpts were presented in random order, reducing the risk of carry-over and order effects, i.e. where the intensity of a music excerpt may effect the perception of intensity in the next. The subjects were encouraged to spend between 15 and 20 minutes doing the survey. To ensure a stable listening environment, the subjects were led through a sound level calibration procedure and instructed to use headphones.

[3] http://wordnet.princeton.edu/

[4] http://www.allmusic.com/

For audio data we selected 150 popular music recordings, spanning a wide range of genres and (as far as we could tell) intensity levels. A 20 seconds long excerpt was selected, 30 seconds from the beginning, in order to get more homogeneous segments. (Avoiding lengthy intros etc.) Each excerpt was additionally audited to reveal undesirable characteristics, such as sudden changes in instrumentation or tempo halfway, and possibly replaced.

The subjects were mainly music students, and music technology students and researchers, from four universities in Spain and Norway, and must therefore be considered as trained listeners.

2.3. Extracting low-level audio features

The following feature extraction functions were implemented in MATLAB®.

2.3.1. Dynamics related descriptors

The dynamics of the sound was estimated for two time resolutions, coarse-grained using Root Mean Square (RMS) calculation and fine-grained using Peak Program Meter (PPM) calculation. Both produce a time-domain envelope in dB. We used histograms to approximate the distribution of dB values in the segment, and described the shape of the distribution in terms of percentiles, centroid, spread etc. Figure 2 shows PPM envelopes and distributions for two signals of different intensity, with the extracted features marked.

The RMS values were computed for 50 ms long, non-overlapping frames. Our PPM implementation is similar to the one described by Zölzer [4]. For stereo signals, the RMS and PPM envelopes were averaged prior to feature extraction. The extracted RMS and PPM histogram descriptors were median, 5^{th} percentile, 95^{th} percentile and 90^{th} inter-percentile range, as well as centroid, spread, skewness and kurtosis (excess).

2.3.2. Spectral descriptors

Spectral descriptors were computed using the Short-Time Fourier Transform (STFT) with a frame length of 23 ms, 50% frame overlap and a Hanning windowing function. Descriptor values for the entire segment were modeled as the mean and variance of the individual frame values. The spectrum magnitudes were averaged prior to feature extraction in the case of stereo signals.

The extracted spectral descriptors were spectral centroid, spread, skewness and kurtosis (excess), as well as spectral flatness.

2.3.3. Long-term loudness estimates

Two estimates of long-term loudness were computed from the audio data.

In studies by Soulodre [2], and Skovenborg and Nielsen [1], the *equivalent sound level* (L_{eq}) measure with the *Revised Low-frequency B-weighting* (RLB) has shown to be a reliable, objective loudness estimate of music and speech. Defined as

$$L_{eq}(\mathrm{RLB}) = 10 \log_{10} \left(\sum_{n=1}^{N} x_W[n]^2 \right) \qquad (1)$$

where $x_W[n]$ is the frequency-weighted digital waveform and N is the total length of the signal, our implementation is similar to the one described in [2].

Skovenborg and Nielsen [1] introduced a new loudness model, LARM, optimized for computation of long-term loudness estimates of non-stationary signals. LARM is based on the asymmetrical low-pass filtering of the PPM, combined with RLB frequency weighting and power mean calculation. It achieved best overall performance in their evaluations when compared to traditional loudness models.

2.4. Modeling intensity in music

With a total of 28 low-level features and up towards 150 instances in the training data set, we risk over-fitting the intensity model. It is therefore necessary to reduce the number of features prior to intensity class modeling, using a suitable feature selection algorithm. *Correlation-based Feature Selection* (CFS) has previously shown to excel for feature selection in the context of percussion instrument sound modeling from low-level audio features.

We have made no assumptions about salient features prior to the analysis. Therefore we would like to have a transparent model of the intensity categories, where the significant features can readily be identified and inspected. *C4.5 decision trees* are suitable for this purpose, as the emergent tree structure provides a clear representation of decision boundaries in the feature space.

Both feature selection and the class modeling described in this section was performed with the WEKA [5] data mining software, in wide-spread use throughout the MIR research community.

[5] http://www.cs.waikato.ac.nz/ml/weka/

Feature subset	Accuracy
RMS	28.4%
PPM	47.8%
Spectral	67.2%
Leq(RLB)	40.3%
LARM	43.3%
Combined CFS	62.7%

Table 1. Performance of various feature subsets.

3. RESULTS AND DISCUSSION

The survey was operational for two weeks and resulted in over 3500 intensity category ratings, an average of approx. 23 ratings per music excerpt. The collected data was filtered according to several constraint in order to remove unreliable and possible erroneous contributions, thereby increasing its quality. Since the subjects go through an adjustment phase as they become familiar with the task, their initial contributions are likely to be "out of tune" with the ones following. The first five ratings of each subject were therefore discarded. Ratings that took less than five seconds or more than 100 seconds were also discarded, as this may indicate a lack of focus and possible fatigue.

Most of the 150 music excerpts display a unimodal distribution of ratings among the five intensity categories. Near half of them converged nicely toward one single category. The intensity model was built on a subset of the most consistently rated excerpts, as these are most likely to represent an objective consensus on perceived intensity in music. Excerpts were accepted into the training data set if they had received more than 10 ratings, and if one category had received between 70% and 100% of the ratings. 67 excerpts were selected this way. The rest of the music excerpts formed the test data set that we used to evaluate the final intensity model.

3.1. Evaluation of feature subsets

Table 1 shows the classification accuracy for C4.5 decision trees induced with various feature subsets corresponding to the ones presented in Section 2.3. The combined CFS subsets comes from applying the CFS algorithm to the complete set of features.

It is evident that the fine-grained PPM sound pressure estimates are more correlated to intensity in music than the coarse-grained RMS estimates (which are only slightly better than chance). The PPM histogram features, which measure dynamics and loudness, are only slightly better than the pure long-term loudness estimates $L_{eq}(\mathrm{RLB})$ and LARM.

We see that the purely spectral features subset scores highest, slightly better than the combined CFS subset, which also includes many spectral features. The increase in ac-

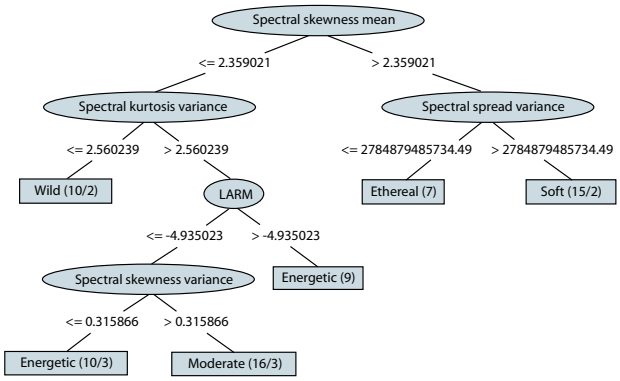

Figure 3. C4.5 representation of the intensity model.

curacy is not large, and since the spectral features subset performs slightly worse on the test data set (Section 3.2), we abandon it in favor of the combined CFS subset.

3.2. Intensity category modeling

A C4.5 decision tree induced with the training data and the optimal feature subset is illustrated in Figure 3. The leaf nodes indicate the predicted category label, the total no. of instances and the no. of incorrectly assigned instances of the training data set.

From the figure we see that the final intensity model is based on only five extracted features, possible evidence of good generalization. The average spectral skewness is a strong salient feature, and good indicator of *ethereal* and *soft* vs. the more intense categories. This reinforces the findings of Zils and Pachet [3], where both the spectral skewness and an EDS variation of it showed high correlation to intensity. The *moderate* and *energetic* categories are not as easily separated as the other categories. This is also the only place where long-term loudness estimate (LARM) appears to be a salient feature.

The confusion matrix resulting from 10-fold cross validation over the training data set is shown in Table 2. The total accuracy of the model is 62.7%, where 20% is the baseline. (Zils' model, built from 18 features, gave an accuracy of 88.7%.) Because of the relatively small data set, this estimate must be interpreted with some caution. We notice that the instances are located very close to the diagonal. This means that, with exception to only three instances, no instances have been misclassified more than one category in either direction, a kind of inaccuracy that may be tolerable in a practical application.

To evaluate the generalizing capabilities of the final intensity model, we applied it to the test data set, i.e. the music excerpts from the survey that were not accepted into the training data set. Since these did not converge towards a single dominant intensity category, we evaluated the means of the subject ratings to the predicted category labels using Pearson's correlation coefficient r. For the test data set we obtained a strong positive correlation of $r = 0.79$, compared to $r = 0.80$ for the training data itself. This shows that the current intensity model is a reliable and reasonably accurate predictor of perceived intensity in music. The intensity model built on purely spectral features obtained a correlation of $r = 0.76$.

True	Predicted				
	Ethr	Soft	Modr	Enrg	Wild
Ethereal	**6**	2	1	0	0
Soft	1	**12**	0	0	0
Moderate	0	3	**8**	4	2
Energetic	0	0	8	**10**	2
Wild	0	0	0	2	**6**

Table 2. Confusion matrix for the C4.5 intensity model.

4. CONCLUSION AND FUTURE WORK

We have presented a descriptive study into perceived intensity of popular music. An intensity taxonomy has been designed for the purpose, and a listening test has been conducted to establish the validity of an objective intensity model. The final intensity model, represented by a decision tree based on low-level audio descriptors, performs reliably for automatic intensity classification.

Intuition tells us that intensity in music is likely to be correlated with mid-level representations such as tempo and "strong beat" (if such), and the effect of adding these to the intensity model will be studied in a forthcoming paper. Furthermore, we must determine the best way to estimate intensity for complete recordings, not just 20 second excerpts. Taking the median of several segments may be a solution.

5. ACKNOWLEDGMENTS

This research was partly funded by the EU project FP6-507142 SIMAC [6] . The first author would like to thank his colleagues at NOTAM for inspiring discussions.

6. REFERENCES

[1] Skovenborg, E. and Nielsen, S. "Evaluation of Different Loudness Models with Music and Speech Material", *Proceedings of the AES 117th Convention*, San Francisco, CA, USA, 2004.

[2] Soulodre, G.A. "Evaluation of Objective Loudness Meters", *Proceedings of the AES 116th Convention*, Berlin, Germany, 2004.

[3] Zils, A. and Pachet, F. "Extracting Automatically the Perceived Intensity of Music Titles", *Proceedings of the 6th Int. Conference of Digital Audio Effects (DAFX-03)*, London, UK, 2003.

[4] Zölzer, U. *DAFX - Digital Audio Effects*. John Wiley & Sons Ltd, England, 2002.

[6] http://www.semanticaudio.org/